THE BELL SYSTEM

Technical Journal

DEVOTED TO THE SCIENTIFIC AND ENGINEERING ASPECTS OF ELECTRICAL COMMUNICATION

VOLUME XXXIV NOVEMBER 1955 NUMBER 6

THE BELL SYSTEM TECHNICAL JOURNAL

THE BELL SYSTEM TECHNICAL JOURNAL is published six times a year by the American Telephone and Telegraph Company, 195 Broadway, New York 7, N. Y. Cleo F. Craig, President; S. Whitney Landon, Secretary; John J. Scanlon, Treasurer. Subscriptions are accepted at $3.00 per year. Single copies are 75 cents each. The foreign postage is 65 cents per year or 11 cents per copy. Printed in U. S. A.

THE BELL SYSTEM
TECHNICAL JOURNAL

VOLUME XXXIV NOVEMBER 1955 NUMBER 6

Silver Migration in Electrical Insulation

By G. T. KOHMAN, H. W. HERMANCE, and G. H. DOWNES

(Manuscript received May 17, 1955)

Silver migration may be defined as a process by which silver, when in contact with insulating materials under electrical potential, is removed ionically from its initial location, and is redeposited as metal at some other location. This process requires adsorption of water on the insulation surface. Silver is unique in this respect in that it is easily oxidized and reduced and does not passivate. Other metals do not present a practical migration hazard. Presented herein are examples of actual experience wherein silver migration caused trouble, and an explanation of this phenomenon based upon chemical and physical considerations and related laboratory evidence. It is concluded that silver should be used with great caution under the conditions noted.

INTRODUCTION

Silver or silver plated metals, because of the favorable electrical and chemical properties of silver, have application in communications systems and throughout the electrical industry. When fabricating apparatus component assemblies, it is frequently expedient to assemble the silver parts in a manner such that they are in intimate contact with insulating materials such as phenol fiber. When standing electrical unipolar potential and atmospheric moisture are present, the silver may migrate. Silver migration may be defined as a process by which silver, when in contact with insulating materials under electrical potential, is removed ionically from its initial location, and is redeposited as metal at some other location. The redeposition occurs in two principal ways, namely;

as dendritic structures, usually growing from a cathodic conductor or area toward the anode or as colloidal deposits, usually in the area near the anode. The result may be lowered insulation resistance or dielectric failure of the insulator in aggravated cases. Insulating materials must be selected with care if this hazard is to be avoided.

PRACTICAL ASPECTS OF SILVER MIGRATION

Silver Contacts in the Step-by-Step System

Let us look at a service application situation in a communications system where silver migration caused trouble. It will serve to illustrate conditions under which migration may be expected to occur and will assist in an understanding of the problem. Silver migration was experienced in the step-by-step dial system and it will suffice to know that in this system a telephone connection is established by electromechanical switches through a series of bank contacts. Silver plating of the tip and ring bank contacts was standard practice for several years and was abandoned when the migration difficulty became evident.

At this stage of the discussion we will concern ourselves only with the service conditions which produced migration. These factors are time, standing unipolar dc potential and atmospheric moisture. Chemical and physical aspects will be discussed hereinafter.

Through Migration

Fig. 1 is a section through part of a step-by-step bank and shows that the "T" and "R" terminals rest against a phenol fiber separator which is about 0.015″ thick. The standing potential is 48 volts dc which remains unipolar in the direction shown except during the time when the contacts are engaged in serving a call. During this time the potential reverses. This interval amounts to approximately 2 or 3 per cent of the total elapsed time, and since the migration process is irreversible, the effect of the potential reversal for this short period of time can be neglected. This particular manifestation is termed 'through migration".

Four years following the service date of the first office to be installed with silver plated contacts, a trouble condition developed in which the switches failed to pulse accurately, thus resulting in wrong numbers. It was found that the phenol fiber separator between the T and R terminals had suffered from silver migration to a degree sufficiently severe to cause dielectric puncture and carbonization of the insulator, thus producing a momentary short circuit or flashover condition during pulsing. Although the standing potential is 48 volts dc the transient peak voltage impressed on the insulator during pulsing is several hundred volts aris-

ing from the abrupt break of an inductive circuit. Chemical analysis verified the presence of silver in the insulator.

Measurements of insulation resistance in this office showed that the insulators were not affected uniformly. In most cases it appeared that the insulators had not degraded to the point of dielectric failure, but they did show an apparent reduction in insulation resistance of varying degrees of severity. Fig. 2 shows the physical appearance of such insulators when stained by silver migration. An interesting aspect of the measurements was that the value of insulation resistance measured before pulsing dropped to approximately one-half of this value immediately following pulsing. After an elapsed time of about 30 minutes the resistance level restored to the value that was measured preceding pulsing.

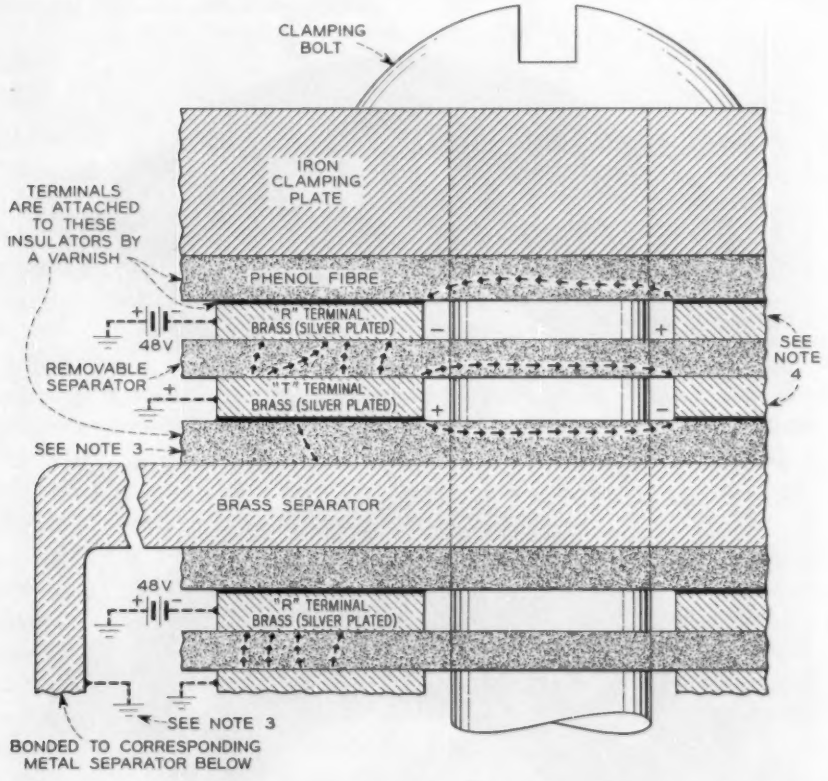


FIG. 1 — Construction of step-by-step bank showing silver migration paths. 1. Arrows indicate direction of migration and the places at which it may occur. 2. Terminals have silver plating on all surfaces. 3. Migration through this insulator can be prevented by grounding the metal separator thus eliminating the potential difference between the tip terminal and the metal separator. 4. Adjacent terminals usually have the same polarity but on connectors serving party lines, they may have opposite polarity and lateral migration of silver may occur.

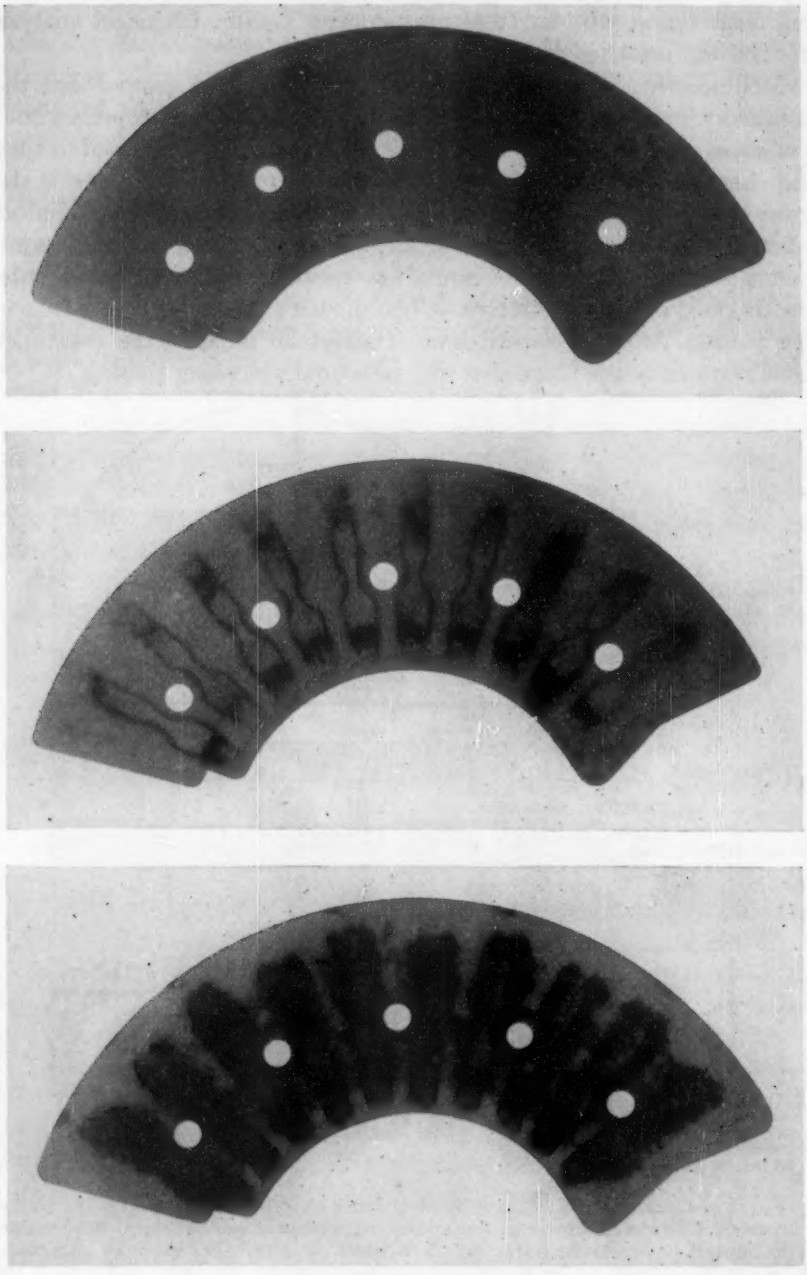


FIG. 2 — Step-by-step bank phenol fiber separators illustrating silver migration.

An equipment unit consists of either 10 or 11 banks and the associated connecting wiring, and this unit is termed a bank multiple. The initial insulation resistance of a bank multiple when new is in the order of 1,000 megohms, and the resistance can degrade to a value of approximately 0.03 megohms in the entire talking path through an office before pulsing failures are likely to occur. By testing banks in the laboratory it was found possible in thirty days, by employing accelerated test conditions of 130 volts at 90 per cent relative humidity, to duplicate the amount of silver transferred into the insulator in four years of service.

There is also a possibility of through migration into the insulator to which the terminals are attached, as indicated by Note 3 of Fig. 1. This possibility can be prevented or stopped by grounding the metal separator, thus eliminating the potential difference between the tip terminal and the metal separator.

Lateral Migration

In the step-by-step bank system there is another migration path, illustrated by Note 4 on Fig. 1, which is termed "lateral migration". This occurs in only a relatively few cases and is most prevalent where the standing potential on laterally adjacent terminals in the idle condition is of opposite polarity. This situation occurs on connector switch banks in the case of certain party lines. This particular manifestation of migration generally takes the form of a deposit of silver, of superficial depth, on the front edge of the insulator to which the terminals are attached. In aggravated instances there may be some penetration back into the insulator. Deposition of silver along the front edge of the insulator can bridge terminals laterally and has caused crosstalk or a trouble condition which involves premature tripping of the ringing on a subscriber line, in which case a "no ring" situation results. Fig. 3 shows an insulator of this type which has failed dielectrically.

The lateral migration condition just described is the usual one with respect to the insulator to which the terminals are attached. However, there is another migration condition which has resulted from a slightly different bank construction. In some of the earlier banks which had silver plated terminals, the terminals were attached to a strip of varnished cambric which rested against a phenol fiber insulator. In this case the silver migrated into the cambric in a manner as shown in Fig. 3.

Silver Charged Dust

There is one further manifestation of silver migration which has occurred very infrequently in the step-by-step system. This involves a

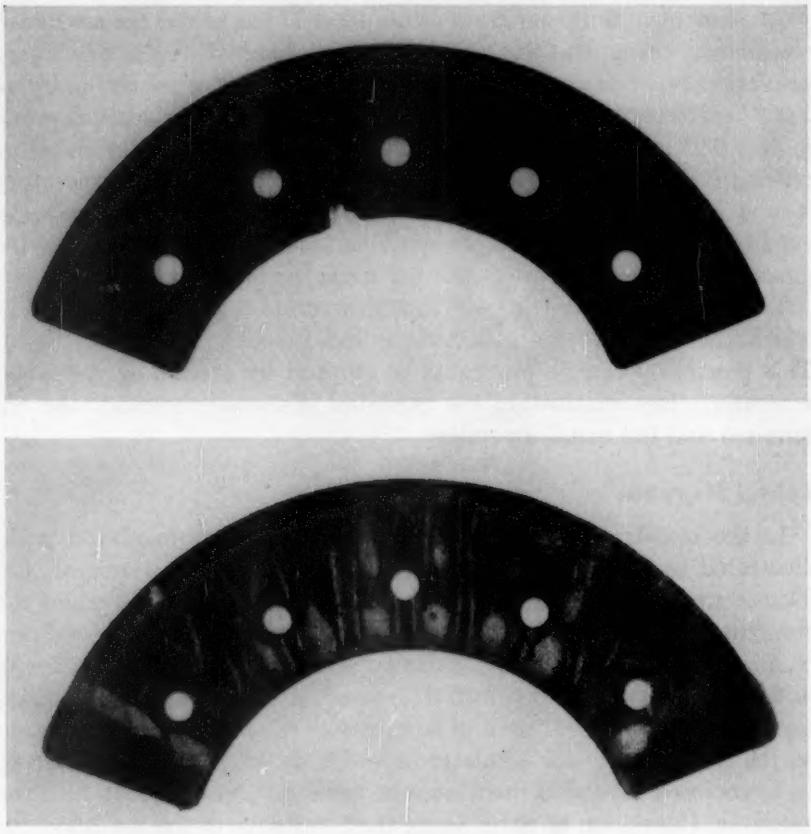


FIG. 3 — Step-by-step system — illustrating migration in insulators to which the terminals are secured.

condition in which substantial quantities of dust have been allowed to accumulate in the bank interspaces between terminals laterally. In this case the dust may become loaded with silver and this can result in sufficient resistance coupling between terminals to produce some crosstalk. Presence of silver in these dusts was verified by chemical analysis.

Panel Dial System

Another application of silver plated bank contacts was in the panel dial system. At one period of manufacture the contacts in the banks, through which talking connections are established, were silver plated in their entirety. Later, the contacts were plated in a manner such that the silver was omitted from the terminal area in contact with the insu-

lation. Still later, following the introduction of silver multiple brushes, it was no longer necessary to plate the associated bank terminals and silver plating of the brass terminals was discontinued. In certain of the frame circuits, standing 48 volt dc potential is present during the idle circuit condition in a manner similar to that of the step-by-step system. The bank construction differs materially from that of the step-by-step system. The panel bank consists of long continuous metallic strips of terminals which are insulated from each other by a wood pulp paper impregnated with asphalt. More recently, some difficulty from silver migration has been experienced in this system as manifested by lowered insulation resistance or momentary short circuits. Silver migration in this instance usually takes the form of very fine silver filaments which grow on the surface of the paper insulation. Bank vibration and dimensional changes in the bank structure, resulting from shifts in the level of relative humidity, may cause the fine silver filaments to break or become reoriented thus producing an unstable electrical leakage condition. This problem is in the investigative stage.

Field Measurements

Measurements by field maintenance personnel in step-by-step offices have generally been made with a 1,000-ohm per volt 100-volt scale voltmeter using 100 volts of test battery. Measurements are thus in terms of volts of leakage which is converted to ohmic resistance. Many readings have been taken which show varying degrees of leakage. In conducting more detailed studies of a field nature, a megohm bridge is used to obtain more accurate resistance data. This instrument measures ohmic resistance directly. Table I lists typical measurements in a step-by-step office, when using a megohm bridge, on banks where migration had not reached an advanced stage.

The readings in Table I show the insulation resistance for a bank multiple and represent lateral migration leakage wherein the minimum distance between edges of the terminals is 0.120″. The silver plated multiples had been in service for 9 years. At the time the measurements

TABLE I — TYPICAL MEASUREMENTS IN A STEP-BY-STEP OFFICE

Measurement Points	Insulation Resistance in Megohms	
	Silver Banks	Brass Banks
Tip to Tip and Tip to Ground	309	1177
Ring to Ring and Ring to Ground	416	983

were taken the temperature was 87°F and the relative humidity 59 per cent.

These data serve to illustrate the gradual deterioration of the insulation resistance of silver plated contact bank multiples. Measurements made in the laboratory on individual contacts show that the insulation resistance tends to remain at a relatively high level until the point of dielectric failure is imminent. This leads to the conclusion that when evaluating the silver migration hazard with respect to insulating materials, the test should be made in terms of both dielectric strength and insulation resistance.

Field Aspects of Laboratory Testing

The principal objectives of the material laboratory testing program were as follows:

1. To verify the fact that silver migration was present.
2. To study the factors which affected the rate of migration.
3. To determine whether temporary remedial measures could be devised.
4. To study materials to determine those which would not be subject to silver migration and which could be used to replace defective insulators in service.

This program was carried to a successful conclusion.

Verification of the fact that silver migration had occurred was simple and conclusive.

The factors which affected the rate of migration were found to be the nature of the insulating material, the magnitude of the standing unipolar dc potential, the elapsed time of standing voltage and the level of relative humidity. Phenol fiber and phenol fabric were found to be particularly susceptible to silver migration. Increasing the voltage and maintaining a high level of relative humidity were found to accelerate migration. Many insulating materials contain salts which, depending upon their nature, may or may not affect the rate of migration. It was found that the relative humidity must be maintained at a very low value to completely inhibit migration. The development of migration in the field was most severe in high humidity areas. Leakage readings made during periods of high humidity were always greater than similar measurements made at low humidities. Such expedients as reversing the dc potential were tried but it was found that a constantly reversing potential at a rapid rate of once per second was necessary before substantial benefit could be realized.

It is fortunate that the progress of migration is rather slow at 48 volts,

and time was available to devise remedial measures. A number of such measures were investigated most of which proved ineffective, and with one exception, none was found which had practical application. The one method which appeared promising was to subject the phenol fiber to heat. While this treatment did not effect a permanent improvement, it did restore insulation characteristics for an appreciable period of time.

Many materials were studied to determine their silver migration susceptibility. A surprising number of these materials reacted unfavorably. Two materials of practical value were found; namely, a special rubber composition sheet material, with the surface sandblasted, and an acetobutyrate plastic sheet material. Neither of these materials was found to be subject to either through or lateral surface silver migration.

FIELD REMEDIAL MEASURES

Heat Treatment of Banks

The first installation in which silver migration was experienced required that remedial measures of some sort be made available promptly if replacement of certain equipments was to be avoided. The only method which could be recommended on short notice, based upon laboratory tests, was to subject the banks to heat. It was predicted that sufficient improvement in the insulation characteristics would result to correct service troubles. Furthermore it was hoped that the improvement would be effective for a sufficient period of time to permit the determination of a more permanent remedy. The heat treatment method was used and in most instances improved the insulation characteristics to a point where service difficulties were no longer experienced and the improvement persisted for a time sufficient to develop a more permanent remedy. Hot air at a temperature of approximately 350°F was forced through each bank for a period of about one-half hour. Fig. 4 shows the heat treatment facilities in place on a bank multiple.

Replacement of Bank Separators

Silver migration in the step-by-step system has affected principally the phenol fiber separator between the T and R terminals. The problem was to discover an insulating material which would not be subject to through migration. The separator may be replaced quite readily by dismantling the bank, removing the degraded insulators and inserting new insulators. This is a simple procedure as contrasted with replacement of the entire bank multiple.

The material selected initially for the replacement insulator was a

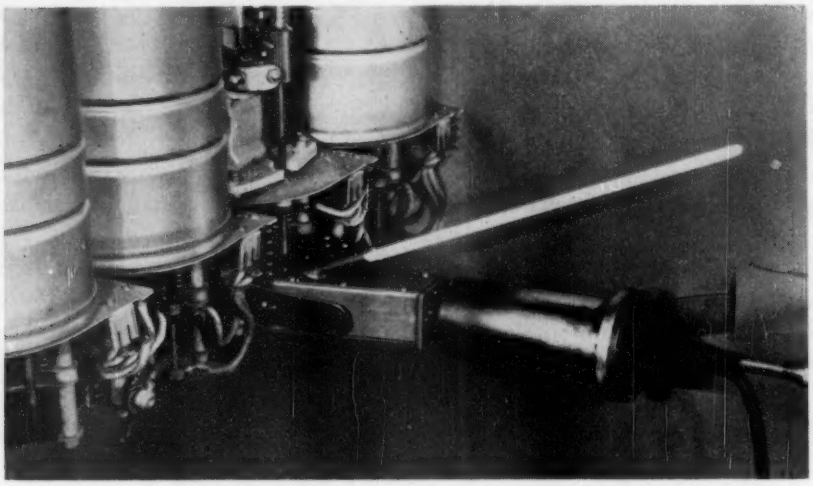


FIG. 4 — Equipment for heat-treating step-by-step banks.

special hard rubber composition with the flat surfaces sandblasted. Insulators of this type have been used successfully and after several years of service have exhibited no silver migration contamination.

For a short period of time insulators of the same rubber composition, with the sandblasting omitted, were tried, but crosstalk troubles developed in service in a relatively short time. While this rubber was not subject to through migration it did develop a lateral migration condition on the surface as illustrated by Fig. 5. It was found that the surface contamination consisted of silver and silver sulphide. It was again necessary to replace these insulators.

By this time a material superior to rubber had been tested and qualified as being free from either lateral or through migration hazards. This material was an aceto-butyrate composition sheet. It has been employed for all subsequent replacements and has proven to be free from silver migration impairment after several years of service.

Lateral Migration

This condition usually manifests itself in the step-by-step system as a continuous deposition of silver of superficial depth on the front edge of the insulators to which the terminals are secured. Correction of this condition involved the development of tools and procedures for physically removing this thin layer of silver. Scraping tools were recommended and used with considerable success for correcting this condition. For those

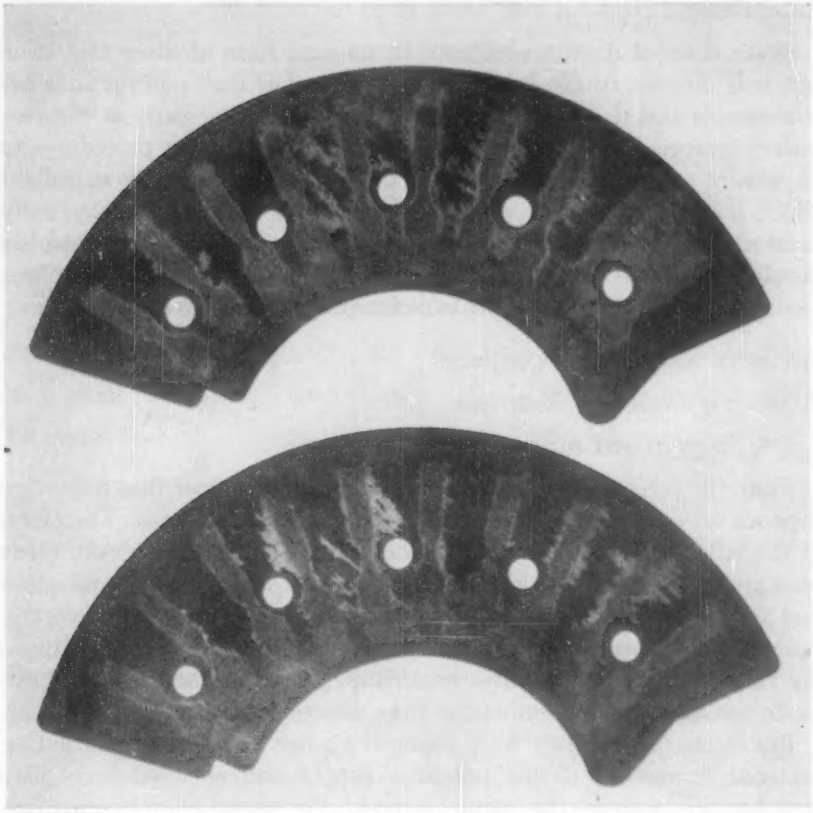


FIG. 5 — Step-by-step bank hard rubber separators illustrating lateral surface migration.

few cases where the silver had penetrated into the insulator to an appreciable distance back from the front edge, the scraping technique was ineffective. For the occasional replacement of such insulators, a set of tools and fixtures was made available to make it possible to remove and replace an individual insulator. In this operation, which is more complex than the replacement of a removable bank separator for through migration, the bank was dismantled, the terminals were broken loose from the phenol fiber insulator and an aceto-butyrate insulator was used for replacement. This procedure would be too costly if many insulators in a bank multiple were affected. Thus in extreme cases of lateral migration affecting a multiplicity of insulators in a bank multiple, replacement of the multiple was the only practical expedient.

Silver Charged Dust

Silver charged dust results from an unusual form of silver migration but it is difficult to combat when it occurs. The dust compactions are inaccessible and they adhere firmly to such insulating parts as the varnished cambric component of certain banks, and cleaning procedures to physically remove the dust cannot be employed. The only expedient which shows promise is to apply a high voltage discharge to physically burn silver compounds out of the compacted dust. Very little need has developed for a remedy for this condition and the burn-out technique has not progressed beyond the experimental laboratory stage.

CHEMICAL AND PHYSICAL ASPECTS

Behavior of Insulating Materials

a — Structure and surface

From the preceding discussions and evidence, it is clear that migration depends on water and on structure of the insulating material. Therefore, in the laboratory investigations of the process, observations were made on a group of materials which differ widely in their affinity for moisture and also in their structure. To establish the possible relation of electrolytic contamination, observations were also made on materials containing varying amounts of soluble impurities. The experimental procedures made use of controlled humidities over various saturated salt solutions.

Bar electrodes of silver were clamped against strips of the insulating material, connected to the potential supply and enclosed in a glass chamber. Fig. 6 shows the clamp assembly. For lateral migration studies, spacings of $\frac{1}{8}''$ to $\frac{1}{2}''$ were used.

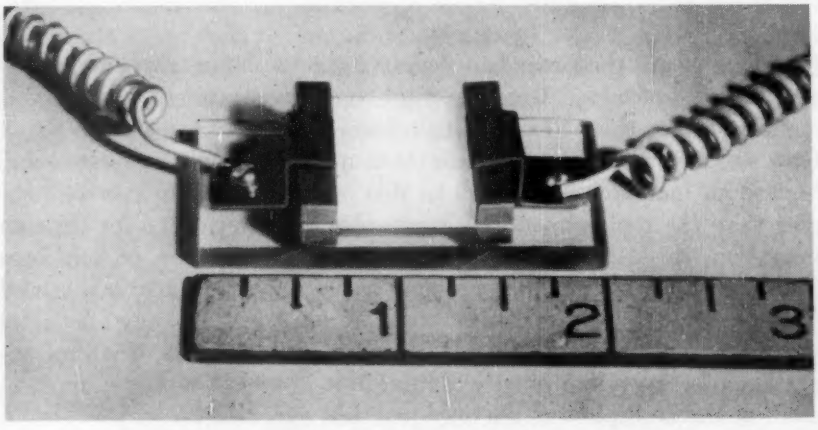


FIG. 6 — Electrode clamping assembly for migration studies.

TABLE II — MIGRATION TEST WITH SILVER ELECTRODES ON VARIOUS MATERIALS AT 91 PER CENT R.H.

Material	Thickness	Applied Voltage	Resistance, Megohms, Final	Length of Test in Hours	Behavior
	inches				
Polystyrene (G.E. No. 1421)	5/32	200	10^5	3,100	No migration
Tenite I	0.010	200	10^5	5,300	Very slight migration
Tenite II	0.018	200	10^5	5,300	No migration
Hard rubber separator sandblasted*	0.019	200	10^6	3,100	No migration
1-Stage unfilled Bakelite resin (No. BR-15,055)	0.075	200	10^6	1,900	No migration
2-Stage unfilled Bakelite resin (No. BR-1,922)	0.065	200	3×10^4	1,900	Very slight migration
Ruby mica	0.002	200	2	3,700	Slow surface migration
1 S phenol fiber	0.019	200	0.01	2,600	Appreciable migration—started after 140 hours
Microscope slide glass	1/32	200	10^5	4,000	Slight surface migration
Absorbent kraft paper	0.006	45	0.01	70	Appreciable migration
White cotton rag paper	0.007	45	0.01	70	Appreciable migration
White a-cellulose wood-pulp paper	0.007	45	0.01	80	Appreciable migration
Whatman filter paper No. 1	0.007	45	0.01	90	Appreciable migration
Lens paper	0.001	45	0.01	48	Appreciable migration
Cellophane	0.001	45	0.15	7	Appreciable migration

* 67.5 per cent Smoked Sheet, 30 per cent Sulfur.

A dc potential of 45 volts was used in most cases. The temperature was maintained at 35°C and frequent measurements of resistance were made. In some instances, where the resistance remained high, the voltage was later increased to 200.

Table II shows the comparison of migration in a variety of materials at 91 per cent R.H. The change in resistance for these materials is shown graphically in Figs. 7 and 8. The order of the materials in Table II is approximately that of their affinity for moisture. As expected, this corresponds very closely to the extent of silver migration. The extreme range of behavior is indicated by comparison of polystyrene, which shows no migration at 200 volts after 3100 hours, retaining a final resistance of 10^5 megohms, with cellophane the resistance of which falls to 0.15 megohms in 7 hours at 45 volts. Tenite I and II show little or no migration under the most severe conditions. Sandblasted hard rubber

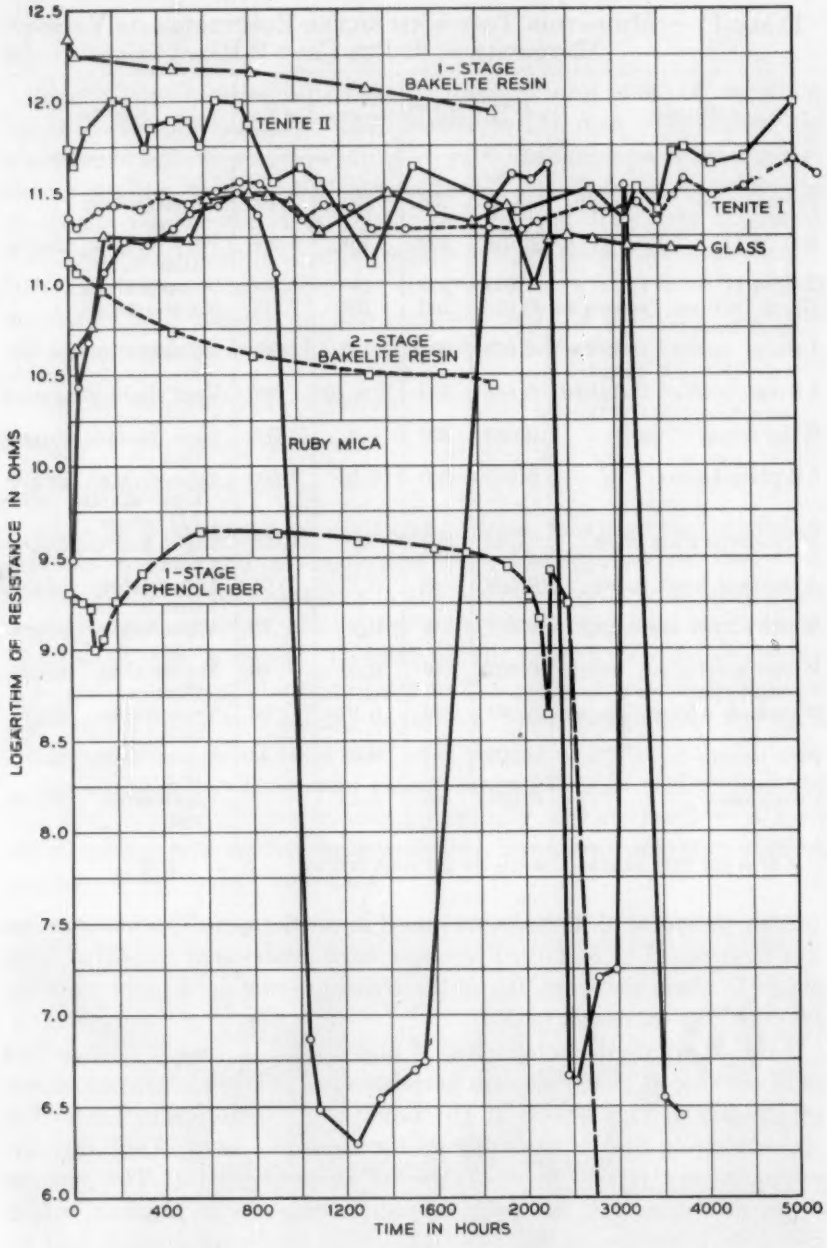


FIG. 7 — Silver migration in various insulating materials.

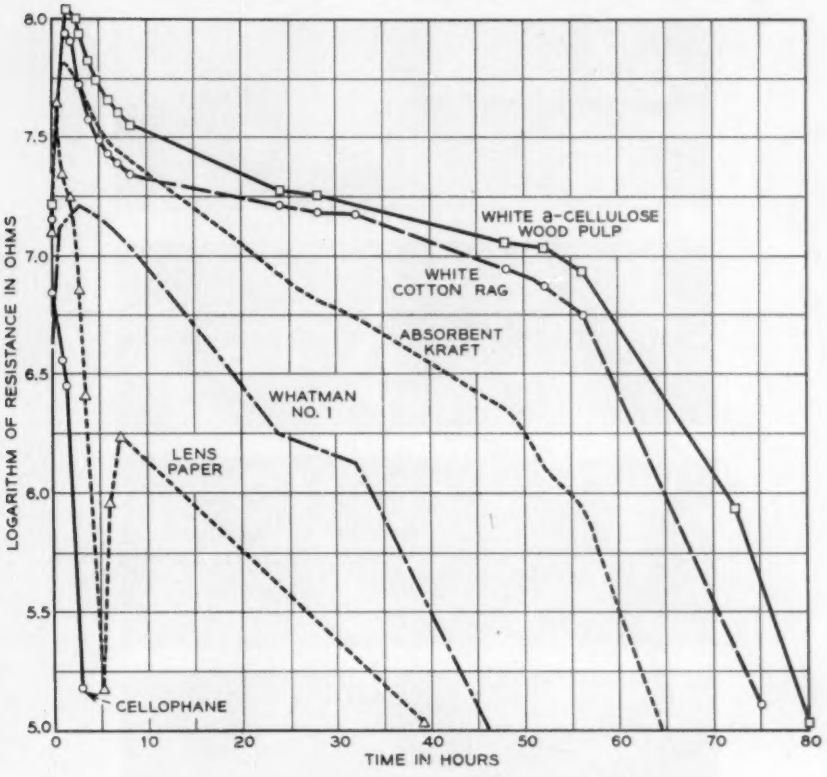


FIG. 8 — Silver migration in cellulosic materials.

which has a somewhat greater affinity for moisture than polystyrene also shows no migration. Slight migration was observed in two-stage unfilled Bakelite resin as compared with none in the single-stage resin. This is probably caused by a slight excess of catalyst usually found in the two-stage material, the decomposition products of which tend to dissolve silver in the presence of moisture. Mica and glass, the surfaces of which are much more hygroscopic, exhibit surface migration visually as well as electrically.

The cellulosic materials which are considerably more hygroscopic than any of the other materials tested also exhibit extreme silver migration. In these cases, because of the porosity of the material, both lateral and through migration are observed. Furthermore, the high purity papers, such as the absorbent Kraft paper and the white cotton rag insulating papers show less migration than the less pure lens paper and cellophane.

The appearance of the specimens after test is shown in Figs. 9 and 10.

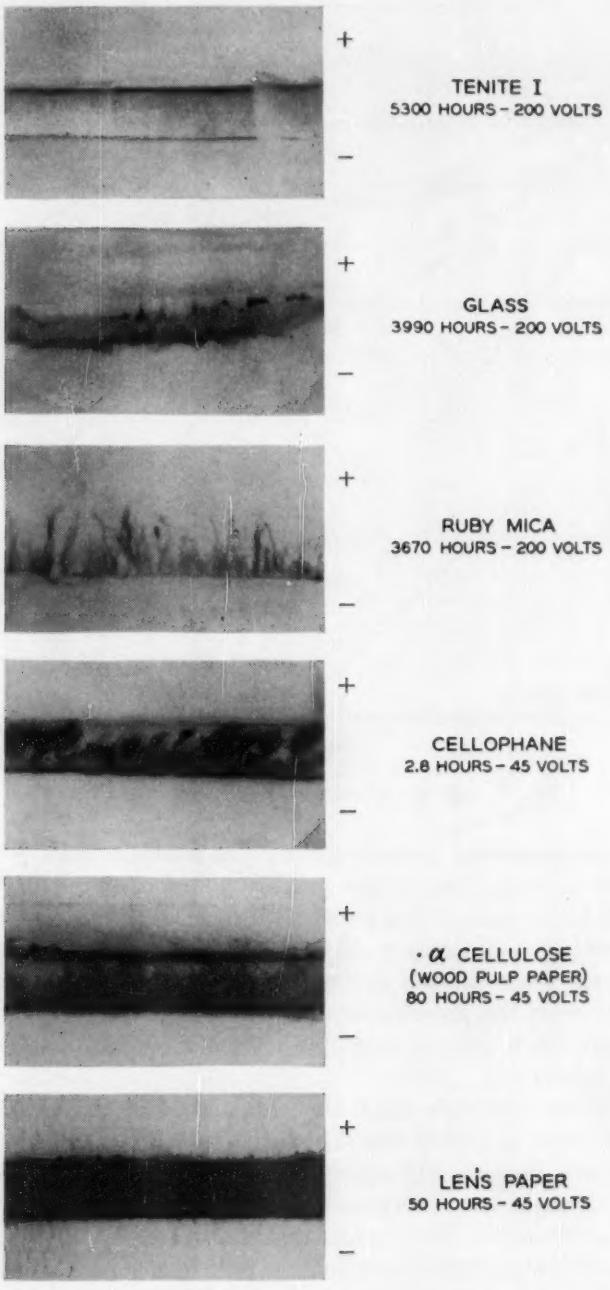


FIG. 9 — Silver migration patterns in various materials at 91 per cent R.H.

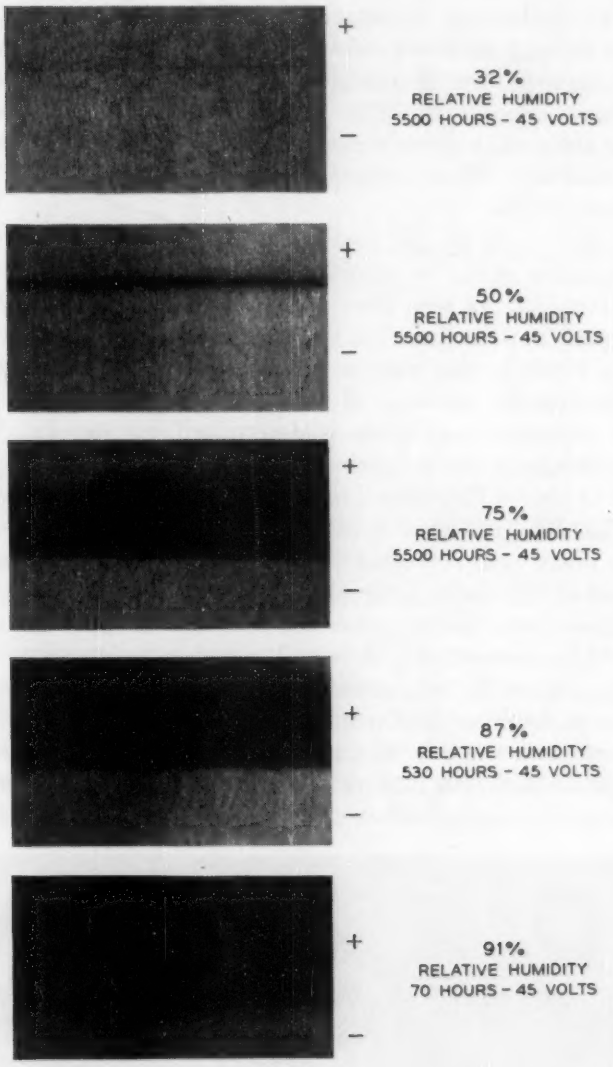


FIG. 10 — Silver migration in kraft paper at 45 volts and various relative humidities.

Microscopic examination reveals a dendritic growth from the cathode (−) and a brown deposit at the anode (+). As shown by Figs. 9 and 10 and the data of Figs. 7 and 8, the deposits grow until they bridge the electrodes causing a sudden drop in resistance which in some cases is followed by a sudden increase indicating partial destruction of the conducting

paths. The mechanism of destruction may involve recrystallization of the silver brought about by resistance heating. It is well known that at moderate temperatures thin conducting silver films recrystallize to form discontinuous deposits. Since the silver is not eliminated by this process the paths are usually quickly reestablished, resulting in another sudden drop in resistance. These experiments were not continued to the point of dielectric failure.

In the case of cellophane, manufacture involves extrusion through a slit. Orientation of the linear cellulose molecules occurs in the direction of this extrusion. As seen from Fig. 9, migration does not follow the shortest path but the direction is determined by the orientation of the molecules which in this case is at 45 degrees from the applied field. This illustrates the tendency of the ions to follow the path of highest moisture concentration which is determined by the distribution of hydroxyl groups in the cellulose molecules.

In the 1S phenol fiber lateral migration is readily apparent after 1,000 hours. The dendritic growth extending from the cathode follows the cellulosic fibers until it reaches the anode. Fig. 11 shows the appearance at the end of this time. Little change in resistance was observed until the electrodes were bridged, when the resistance dropped abruptly. This appears to be characteristic of lateral migration in impregnated fibrous materials, and for this reason resistance measurements are not a reliable indication of the progress of migration.

Comparison of the unimpregnated fibrous base (paper, for example) with the same materials into which resins have been introduced (phenol fiber) shows the marked effect of the latter in reducing the rate of mi-

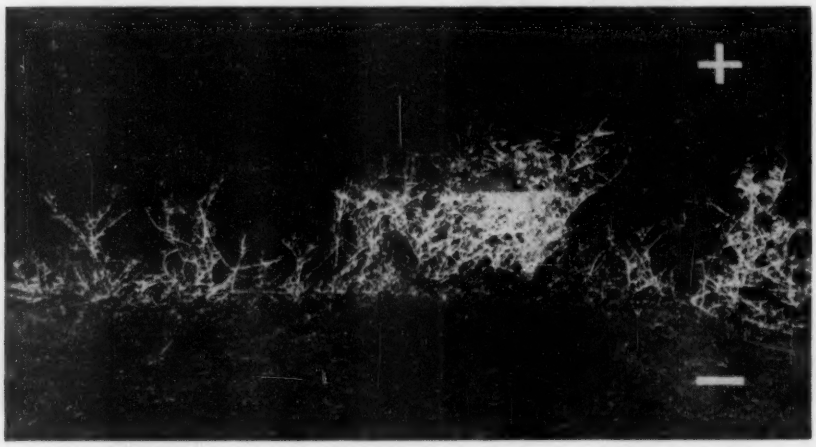


FIG. 11 — Lateral migration on 1S phenol fiber after 1,000 hours, 91 per cent R.H. photographed by reflected light to show silver dendrites.

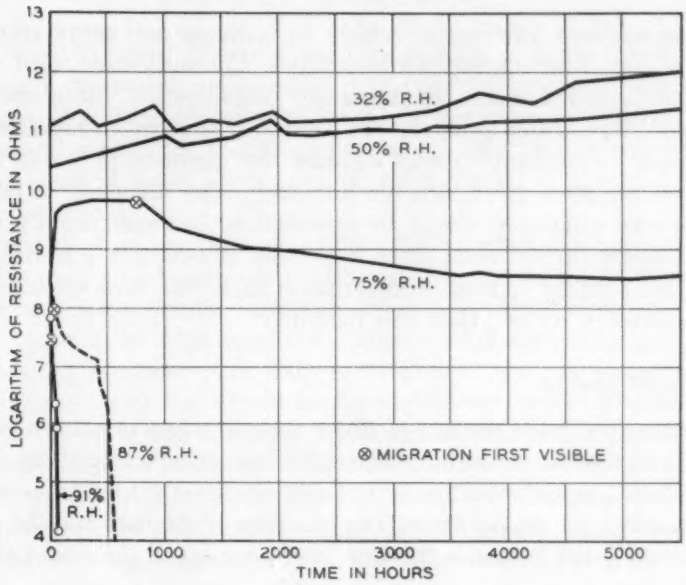


FIG. 12 — Resistance of kraft paper as related to humidity.

gration. Such impregnants restrict ion mobility, probably in two ways: (1) They reduce the rate of moisture penetration, thus providing protection against transient high humidities. (2) They interpose barriers between the individual fibers. From the practical standpoint, however, such improvement is still not effective enough as evidenced by the occurrence of silver migration in phenol fiber in the field.

b — Moisture

As has been mentioned, the affinity of surfaces for water varies widely. The surfaces of non-polar materials such as polystyrene and polyethylene have only a slight attraction for water which is reflected in a high angle of contact between liquid water and the solid surface and also by a low heat of adsorption. On the other hand, the polar hydroxyl group in cellulosic materials gives them a very high affinity for moisture. The dependence of silver migration in Kraft paper on relative humidity is shown in Fig. 12. The insulating materials promote silver migration primarily by providing surfaces on which it is possible for moisture to liquefy at relative humidities below 100 per cent. It is very probable that for silver ions to form and move on a surface, at least a monolayer and possibly two must be present.* To cause this amount of water to con-

* For a discussion of mobility in adsorbed layer see Statistical Thermodynamics, Fowler and Guggenheim, Cambridge University Press, 1939, page 421.

dense on surfaces differing as widely as cellulose and polystyrene requires widely different relative humidities. Contamination with water soluble compounds is also an important consideration. Since contaminants and degradation products are inevitably present on cellulose surfaces, such contamination may increase the quantity of water which will condense at a given relative humidity. The influence of relative humidity on migration would be expected to decrease rapidly below approximately 30 per cent, since less water is normally adsorbed and since the solubility of most compounds is such that they do not cause condensation of water below this humidity.

Microscopic Studies

Microscopically the pattern of silver deposit, when migration occurs, will vary depending on the structure of the insulating material, its chemical composition and whether or not the electrical field acts *across the surface* or *through* the insulator. The intensity of the field and the available moisture will influence the rate, the manner and the extent of deposition.

When the insulator has an impregnated fibrous structure such as phenol fiber, phenol fabric or asphalt impregnated paper, migration progresses along the cellulose-impregnant interface. In typical manifestations, fine filaments as small as 2 to 5 microns in diameter grow out from the cathode along the individual fiber surfaces toward the silver anode. Eventually these filaments reach the anodic area and complete tenuous metallic paths. Resistance measurements have been made on areas where a multiplicity of such paths exists. In the final stages, when potential is present, the resistance is very erratic, falling, then rising sharply in a matter of seconds. This behavior suggests a continual making and breaking of the filamentous contacts on a micro-scale.

The paths from cathode to anode may be relatively direct in the case of materials in which the fibers are arranged in an orderly structure. Such materials are the impregnated fabrics (phenol fabric, varnished cambric, etc.). In these the impregnant rarely if ever reaches completely into the twisted fibrous bundles forming the threads, and almost straight line paths thus are provided. Materials having the fibers in random orientation such as paper or felt, on the other hand, may provide much longer, more tortuous paths from cathode to anode. In such felted materials, a path is formed only by chance overlapping or crossing fibers with no resin barrier between them. Eventually a circuitous chain is completed from cathode to anode.

Where migration occurs *through* an impregnated paper, the paths are

even more tortuous. Phenol fiber, for example, is made up of several layers of Kraft paper, impregnated with phenolic resin and cured under pressure. The fibers are oriented roughly in the plane of the paper lamination and a path for migration normal to the plane is afforded only when a chain of fibers meets in unimpaired contact at each of the lamination interfaces.

When thin specimens of phenol fiber exhibiting migration are examined microscopically by transmitted light, the structure of the silver deposits can be seen fairly clearly. In the anodic area the fibers have a uniformly blackened, non-lustrous appearance suggestive of deposition of silver by chemical or photochemical reduction. This is not unreasonable since reducing agents such as aldehydes are very likely present in small quantities in the resin and the paper itself may have reducing properties. In advanced stages, the blackness seems to diffuse out from the fibers to invade the resinous material. These colloidal deposits may end close to the anode or they may extend practically to the cathode, depending on the particular specimen of phenol fiber, the conditions of transmitted light, voltage and duration of the migration.

The fibers in contact with the cathode, on the other hand, show silver having a quite different structure. Here the deposits have a decided metallic lustre and follow the fiber-resin interface without diffusion into the latter. Under 200 to 400 magnifications, it can be seen that the silver consists of a fine network of metal growing out from the cathode mostly on the face of the fiber but occasionally found growing in its central canal. Where fibers cross, the silver dendrites often change their course and follow a new fiber, the process being repeated many times as the metal grows out into the available ion paths. The net result is a tree-like macro structure made up of dendritic micro structures on the individual fibers. In Fig. 13 are shown individual fibers with the fine dendrites of silver growing on them. These fibers were obtained from a panel system bank insulator in service. The asphalt impregnant was removed by extraction and the paper teased apart in a suspending fluid. Fig. 14 shows a single cotton fiber which was stretched between silver anode and cathode wires in 98 per cent humidity at 45 volts. After 36 hours a cathodic outgrowth of silver was formed in the canal of the fiber. The anodic end of this fiber showed only a structureless transparent brownish stain.

Electrochemical Mechanism

The experiments and observations to this point confirm the reproducibility of migration in or on a wide variety of materials, the only requirements being a standing potential and high humidity. In order to obtain

FIG. 13 — Paper fibers removed from cathodic area showing dendritic silver deposits.

a better understanding of the nature of silver migration, a series of experiments was carried out with filter paper which was relatively free of impurities. In this way, electrolytes, reducing agents and other complicating substances present in commercial materials are largely eliminated. The absence of impregnating resins also facilitates observation of the process both microscopically and microchemically.

CS&S No. 598 paper was clamped between silver bar electrodes at

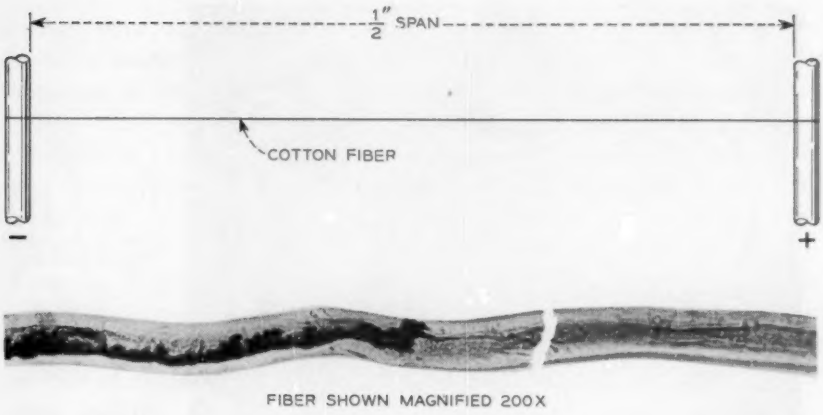


FIG. 14 — Single cotton fiber stretched between anodic and cathodic silver wires showing cathodic silver deposits in central canal, 45 volts, 98 per cent R.H., 36 hours.

98 per cent relative humidity. The spacing between the electrodes was ½″, across which a 45 volt potential was maintained. Fig. 15 shows the condition after 6 hours. Dendrites of silver have grown out from the cathode about halfway to the anode. Outward from the anode is a stained area having a relatively sharp boundary. When first observed, this area is yellowish brown in color but in daylight it changes to a purplish gray, suggesting photo-reduction of a silver compound.

A simple explanation which takes into account these basic observations is the following:

In the electrical field provided, silver ions tend to leave the anode in the water film adsorbed on the cellulose fibers. Hydrogen ions collect around the cathode, where initially their discharge maintains electrical balance. Hydroxyl ions move toward the anode but encounter the silver ions moving from it. A region occurs close to the anode in which the product of the concentrations of silver and hydroxyl ions reaches the solubility product for silver hydroxide and a colloidal precipitate is formed. The unstable silver hydroxide probably decomposes immediately to silver oxide, Ag_2O, producing the dark, structureless deposit already described, around the anode. Subsequently the silver oxide may be reduced by light or by reducing agents present in commercial insulating materials or even slowly by the cellulose itself.

Thus the concentration of the silver ions leaving the zone of Ag_2O precipitation will be regulated by the equilibrium:

$$Ag_2O + H_2O \rightleftarrows 2\ AgOH \rightleftarrows 2\ Ag^+ + 2\ OH^-$$

The solubility product of silver hydroxide ($Ag^+ \times OH^- = K_{AgOH}$)

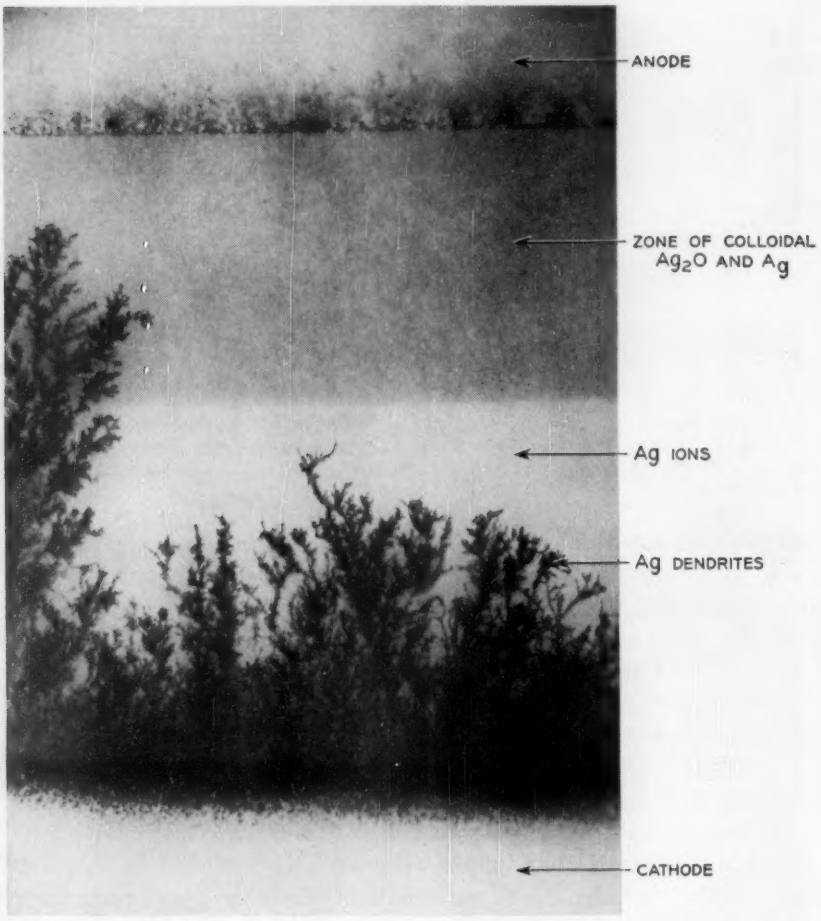


FIG. 15 — Silver migration in CS&S No. 598 paper in 98 per cent R.H., 45 volts. Specimen impregnated with Canada Balsam to Increase transparency.

is 1.5×10^{-8} at 20°C. Assuming, for pure water, that the OH^- and Ag^+ ions are equal, the concentration of silver ions will be $\sqrt{1.5 \times 10^{-8}}$, or about 1.2×10^{-4} moles per liter or about 13 milligrams per liter. Silver ions are released continuously in this concentration and travel on to the cathode where they are discharged with the formation of the characteristic dendritic outgrowths of the metal.

Basically the state of an ionized molecule in solution is not unlike that of an ionized gas, the principal difference being the source of the field which causes ionization. In a water solution the field around the polar water molecule supplies the ionizing potential, while in a gas an external

voltage is required. For this reason the ionization process can be observed in a solution at applied potentials of 0 to 2 volts compared with thousands of volts required for field ionization in gases. In some respects because of the ease of reduction of the silver ion, a solution of a silver salt behaves like a gas near its corona point. In fact there is a strong resemblance between the ionization patterns around a high voltage conductor and the dendritic silver pattern. Once the silver ion has been deposited, the increased field concentration causes other ions to discharge on it preferentially, producing the dendritic growth. In plating metals on electrodes, it is known that the formation of dendrites is favored by a low activation energy of ion discharge at the cathode. In the case of silver, since this activation energy is very low, dendrites are easily formed.

This appears to be a reasonable explanation of the mechanism of silver migration. Wide variations in behavior and pattern are observed under different conditions. It is instructive, for example, to observe the effects of the introduction of various electrolytes into the paper. When pre-impregnated with 0.01M KNO_3, dried and exposed for 16 hours to 98 per cent R.H. at 45 volts, the dendritic growths from the cathode do not appear. In fact there is no evidence that silver ions reach the cathode at all. Rather, they are precipitated, probably initially as the hydroxide at the boundary of the strongly alkaline zone surrounding the cathode. This zone is produced by the local accumulation of potassium ions at the cathode. Hydrogen ions are discharged in their stead and potassium hydroxide results from the secondary electrode reaction. If ammonium nitrate is used instead of potassium nitrate, only a weak alkalinity results, with a much lower hydroxyl ion concentration. Furthermore, the ammonium ion forms the soluble ammonio complex with the silver ions which carries them through to the cathode. Here the complex ion is discharged and the silver dendrites grow outward as in the case of the experiment with water alone. As might be expected, potassium iodide when present in sufficient amount, exhibits a pronounced inhibiting action on the migration because of the insolubility of silver iodide, which precipitates in a zone close to the anode. Making the paper alkaline with potassium carbonate also reduced the migration by favoring retention of the silver as the oxide near the anode.

The question naturally arises as to why, in practice, silver alone seems to be especially subject to the effects of migration. Fig. 16 shows the current-time curves for several metals used as electrodes against moist filter paper. These curves were obtained by using 2″ square electrodes with a pad of six sheets of pre-moistened CS&S No. 598 paper sandwiched

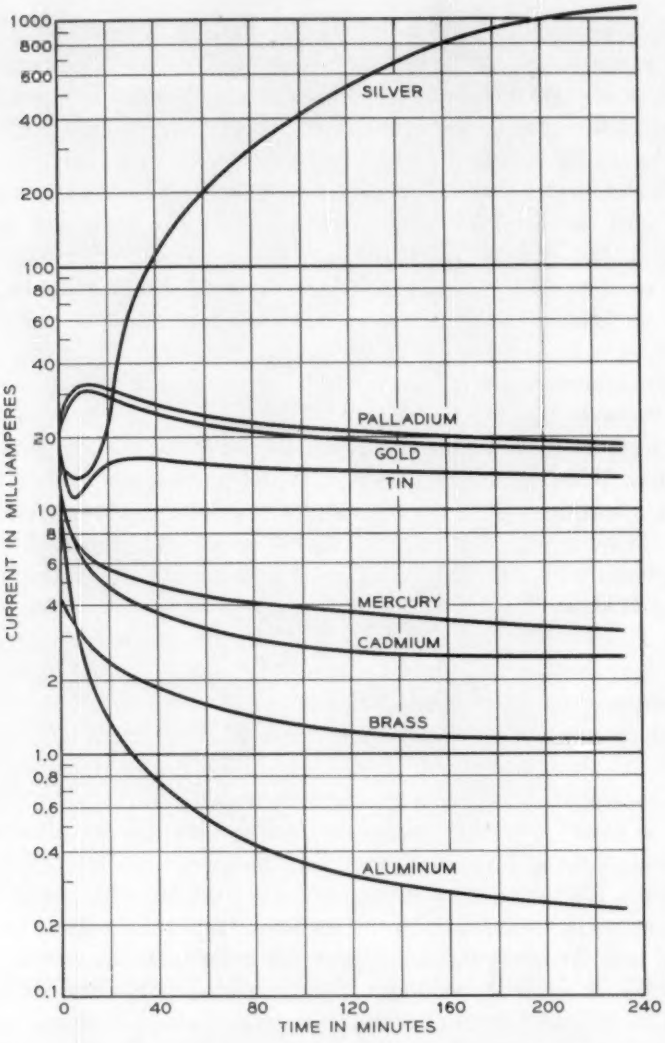


FIG. 16 — Current-time curves for various metal electrodes using moist filter paper, 45 volts.

between. The sandwich was placed under 40,000 pounds pressure to establish good reproducible contact and a 45 volt potential was maintained. In this way sufficient current flow was obtained for direct measurement directly with a milliammeter.

The initial current flow was between 10 and 30 milliamperes for all the metals. For gold and palladium, this level held fairly constant

throughout the experiment. For the base metals (brass, Cd, Al, Hg, Sn) the current dropped off rapidly to lower levels. Silver alone was characterized by a continual rise in the current to a level well above one ampere, at which point considerable heating was apparent.

In each case, the paper sheet next to the anode was removed and tested for the anode metal. The base metals all gave indication of electrolytic transfer, but this was confined to the topmost paper layers in direct contact with the anode and amounted to only a few micrograms. The gold and palladium showed no detectable transfer.

In the case of the silver, the whole pad was filled with reduced metal, from the cathode sheet to the anode.

Thus, it would appear that electrolytic solution of the base metals does occur in the presence of moisture only, but the process is quickly arrested by the formation of highly insoluble, passivating films which seal off the metal effectively. When the paper is removed, part of this film is separated with it and can be detected with sensitive reagents. Gold and palladium having higher oxidation potentials do not dissolve and no passivating films are formed, but rather, oxygen is liberated at the anode, the process producing a fairly constant, low level of current flow.

Silver, on the other hand, dissolves anodically in the presence of moisture alone but is incapable of forming an effective passivating film under such conditions. The oxide formed at the anode is relatively soluble and does not impair greatly the movement of silver ions.

The unique migration behavior of silver, just as its unique value in photography, is probably related to the low free energy of oxidation and reduction reactions. The free energy of formation of the oxide is only $-2{,}395$ calories compared with $-26{,}000$ for copper oxide. Thus silver is readily dissolved anodically but the resulting ions are easily reduced to the metal, both cathodically and by chemical agents.

Inhibiting Techniques

In addition to the remedial approaches already discussed for equipments in service, there remains the question of improving the properties of insulating materials to resist silver migration in new equipment design. In materials having a fibrous base, three approaches are suggested:

(1) Improvement of resin impregnation to isolate fibers more effectively, thus preventing through paths. From the study of the paths observed in actual migration, it seems quite evident that a more thorough impregnation of the paper, especially at the surfaces and at the inter-

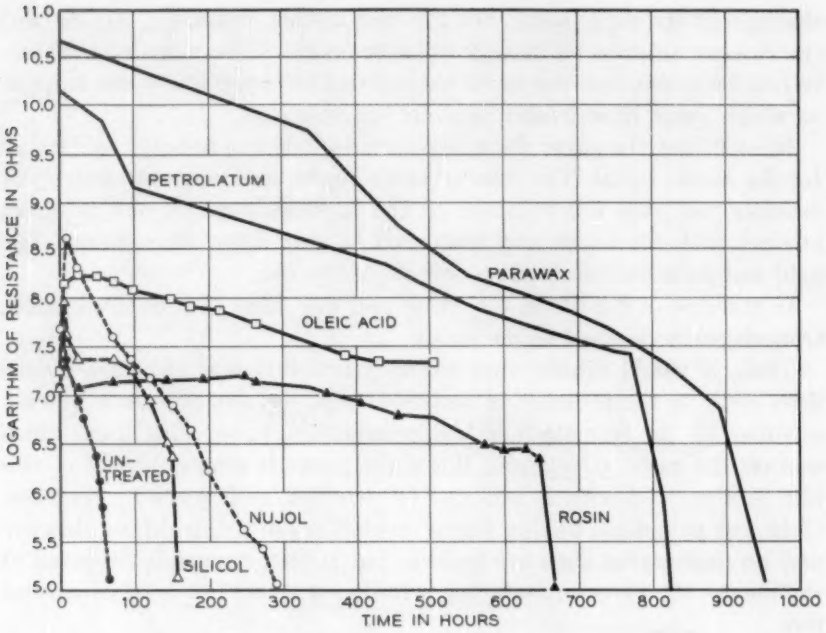


FIG. 17 — Effectiveness of various impregnation treatments in preventing silver migration.

faces between the individual sheets of laminated materials, would greatly reduce the fibrous paths available for migration. Interleaving with a thin sheet of a suitable impervious plastic might provide an effective barrier despite the existence of the fibrous paths on either side.

(2) Treatment to prevent adsorption of water. Pretreatment of fibrous surfaces with water-repellant agents might be expected to reduce silver migration. On the whole, while some reduction was obtained, the results were disappointing. Fatty, waxy and resinous materials as well as silicones and chlorosilanes were tried with varying degrees of retardation but none was sufficiently effective to be practical. Fig. 17 summarizes the performance of a number of these treatments.

(3) Incorporation of silver precipitating agents in the insulation. If a reagent could be introduced into the insulator which would capture the silver ions by precipitation, reduction or sequestration, migration might be retarded to a practical degree. Such a reagent itself would have to be practically non-conducting and its effectiveness would depend on how much of it could be introduced in the migration path. Reducing agents such as aldehydes, hydroquinone, pyrogallol, and hydrazine,

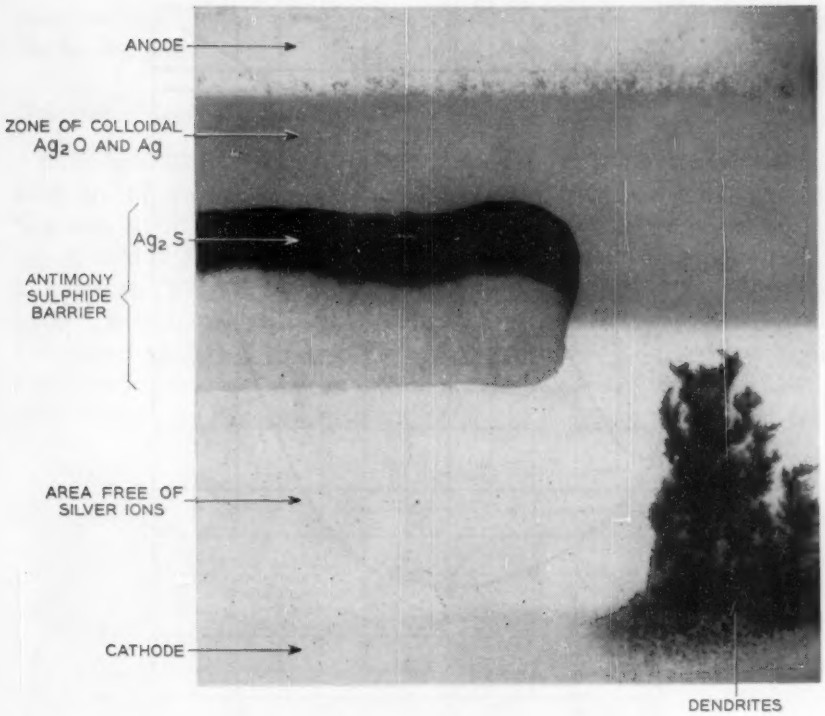


FIG. 18 — Silver migration in CS&S paper with antimony sulfide as barrier, 98 per cent R.H., 45 volts.

when introduced into paper cause the silver to precipitate in the colloidal form in the immediate vicinity of the anode, no dendrites appearing at the cathode until the reducing agent is exhausted. The higher organic acids such as stearic, oleic, and palmitic, appear to have some retarding effect through formation of the almost insoluble silver salts. Benzidine and even egg albumin ties up the silver ions through formation of insoluble complexes. Perhaps the most promising of the precipitating agents, however, is antimony trisulphide, Sb_2S_3. When properly introduced, this compound is quite inert to all but silver and a few other heavy metal ions. It is very insoluble and adds very little conductivity to the paper, yet it precipitates silver ion quantitatively as silver sulphide.

Fig. 18 shows graphically the action of the antimony sulphide. A piece of CS&S No. 598 paper was subjected to silver migration at 45 volts, 98 per cent R.H. Across about two thirds of the paper, a band of antimony sulphide was introduced. Where the silver ions encountered this

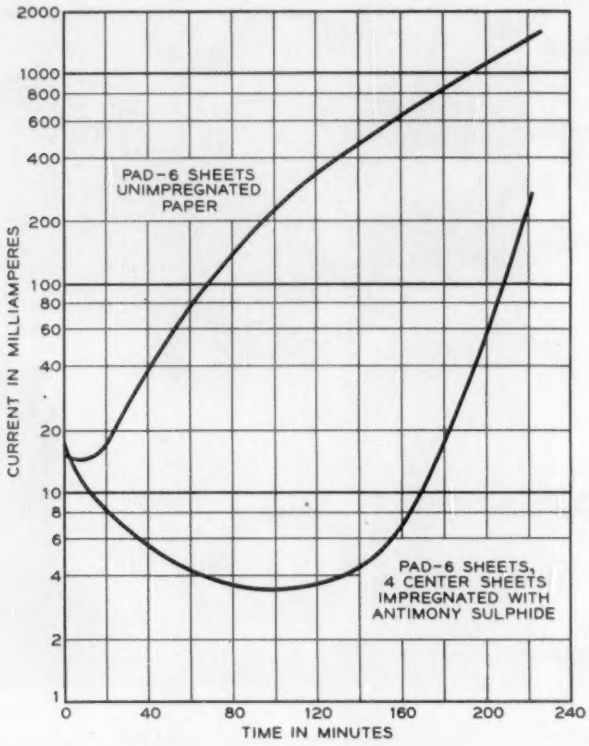


FIG. 19 — Comparison of antimony sulfide impregnated and unimpregnated paper pads.

zone, black silver sulphide was precipitated, none escaping through to the cathode, hence no dendritic growth opposite the band. Where the band is not interposed, the silver ions move freely in the field to give rise to the usual dendritic outgrowth from the cathode. Antimony sulphide is introduced into the paper as the soluble sodium sulphantimoniate. From this, antimony sulphide is precipitated by immersing the dried paper in dilute acetic acid, then the paper is washed free of sodium acetate in running, distilled water.

Fig. 19 shows current-time curves for a 4 square inch pad containing six CS&S No. 598 sheets, the four center ones heavily impregnated with Sb_2S_3. These were first moistened, then placed under 40,000 pounds pressure between the silver electrode platens. A comparison run was made using six unimpregnated papers. This is, of course, a highly accelerated test in which conditions were chosen to favor maximum silver migration. The retardation by the Sb_2S_3, is clearly evident and under

practical conditions it could probably be made to increase the insulation life several fold.

Behavior of Inorganic Surfaces

Brief mention has been made of silver migration on materials such as glass and mica. Evidence has been given in Fig. 7 that surface migration can occur on these materials at high humidities. Ceramics, which are relatively free from water soluble impurities such as high purity alumina and steatite and which possess low surface leakage at high humidity, exhibit migration only at humidites near 100 per cent. It was considered desirable to learn whether migration would occur on a highly insoluble, inert surface such as quartz. A fused quartz plate lightly sandblasted was clamped between silver electrodes in 98–100 per cent

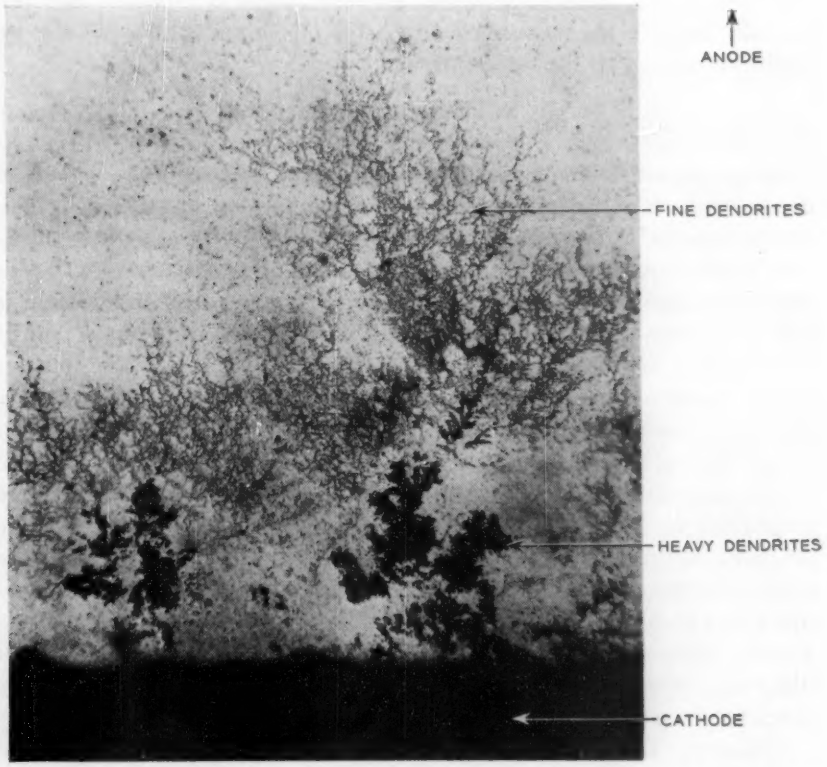


FIG. 20 — Silver migration on quartz plate. Photographed by transmitted light, 100 magnifications, 1500 hours, 98 per cent R.H., 45 volts, ½" interelectrode spacing.

relative humidity at 45 volts. After 1,500 hours, dendritic growths were just visible at the cathode to the unaided eye. The anodic area was practically clear except at the points of contact where a slight brownish stain was visible. Fig. 20 shows an area from the cathode outward, viewed in transmitted light at about 100 magnifications. Short, heavy dendrites are seen growing out of the cathode, after which the paths fan out to form a delicate web reaching toward the anode. This experiment, more than any other, tends to show that the only requirement for the migration, given silver and potential, is an adsorbed water layer. Some concern has been expressed regarding the possibility of migration in silvered mica capacitors. Laboratory tests of the open capacitor indicate no migration over the mica when the relative humidity is held below approximately 1 per cent and only slight migration occurs at 90 per cent. When enclosed in a Bakelite housing, however, no migration was observed under these conditions. There is some indication that the Bakelite resin in the assembled unit gives off vapors which exercise an inhibiting action on the mica surface.

SUMMARY AND CONCLUSIONS

Silver migration is a possibility which must be considered whenever silver conductors, under dc potential, are separated by insulating materials capable of adsorbing moisture. Therefore, silver should be used with great caution when the conditions discussed herein are present and then only after careful testing to insure that the specific application is safe. In porous or fibrous materials, migration tends to follow the pore or fiber surface and may occur in paths *through* such materials. In nonporous materials, migration is restricted to the surface. The process is primarily electrolytic, involving anodic solution of silver in adsorbed water films with formation of silver oxide in the anodic area. The oxide is non-passivating and appreciably soluble so that silver ions are free to migrate to the cathode. There they are discharged to form metallic dendritic outgrowths which eventually may bridge the anode-cathode space. Electrolytic conduction then gives way to metallic conduction and insulation breakdown may occur. In addition to the cathodic reduction process, silver deposits of a colloidal character may result from chemical reduction either of the oxide in the anode area or of silver ions at any point between anode and cathode.

Ordinarily the colloidal deposits account for little of the observed conductivity, the metallic dendrites constituting the main leakage paths. Where migration is confined to a surface, however, and chemical or

photochemical reduction is intense, the colloidal deposit may attain a density sufficient to conduct appreciably.

The combination of properties leading to migration appear to be found only in silver. There is no evidence that migration occurs with other metals under practical conditions.

No completely effective treatment for silver migration has been devised for application to affected equipment in service outside of replacement of the insulator, where this is possible.

As to the improvement of materials for new equipment, there is reason to expect that something can be accomplished, now that the character of the migration process is known. One possibility is the provision of more effective resin barriers in impregnated fibrous materials. Another possibility is the incorporation of agents which reduce or precipitate the silver ions before these reach the cathodic area.

ACKNOWLEDGMENTS

The problem of silver migration has received attention from a considerable number of people in the Laboratories, the A. T. & T. Co., and the Operating Companies during the past fifteen years. A background of experience and opinion thus has been built up, from which the authors have drawn freely. Specifically, the authors wish to acknowledge their indebtedness to J. M. Wilson for his numerous contributions to the field and laboratory evaluation of the migration process and to J. M. A. de Bruyne for his laboratory studies of migration in controlled humidity.

The Field Effect Transistor

By G. C. DACEY and I. M. ROSS

(Manuscript received May 31, 1955)

Previous work on field-effect transistors considered the performance of the device when operated with electric fields in the channel below the critical field, E_c, where the mobility of carriers becomes dependent on field. This work is reviewed and it is shown that, in this range of operation, the frequency cut-off, f, and transconductance, g_m, of the device increase with increasing values of electric field.

New theory is derived for the performance with electric fields greater than E_c, where the mobility is proportional to $E^{-1/2}$. It is shown that, although both f and g_m continue to increase with electric field in this range, the corresponding increase in the power dissipated is so rapid that such designs are unattractive. It is concluded that a good compromise is to operate with the average channel field equal to E_c. The performance in this particular case is considered in detail and the results summarized in a design nomograph. It is found that f is inversely proportional to the "pinch-off" voltage. The pinch-off voltage cannot, however, be made indefinitely small because the gate junction must be in the saturated condition. A reasonable estimate of the minimum voltage is ½ volt and this leads to a maximum value of f of 1,000 mc/s.

A description is given of the fabrication and performance of several field-effect transistors operating in both the constant and non-constant mobility ranges. It is shown that the performance of these units is in agreement with theory. One of these units had a frequency cut-off of 50 mc/s with a transconductance of 1.6 ma/v when operated at 40 volts and 40 ma.

1. INTRODUCTION

Two papers have already appeared on the field-effect transistor,[1, 2] in which the basic features of the theory were presented and their experimental verification discussed. This present paper describes certain additions to the theory and the experimental verification thereof which arise when attempts are made to realize the ultimate in high-frequency performance. In particular the effect of non-constant mobility at high

electric fields is analyzed in some detail and shown to be a governing consideration for some designs.

An attempt has been made to make the present paper sufficiently complete that it will not be necessary to refer to the previously published literature. For that reason certain of the previously published results are quoted as a starting point for the extended theory. The paper is divided into two parts. The first part presents the complete theory as of its present state of development, and the second part describes the most recent advances in the experimental verification of this theory.

Part I, Theory

2. A QUALITATIVE THEORY

In essence, a field-effect transistor can be regarded as a structure containing a semi-conducting current path, the conductivity of which is modulated by the application of a transverse electric field. In particular consider the structure shown in Fig. 1. The device as shown consists of a slab of n-type semi-conductor with an ohmic contact at each end, and two p-type contacts on opposite sides. These p-type regions are called the "gates". Consider what happens if the gates are shorted to the left hand end of the n-type slab and a positive potential, V_0, applied to the right hand end. A current, I_0, will flow between the ohmic contacts and will consist of a flow of electrons from left to right. The left hand contact

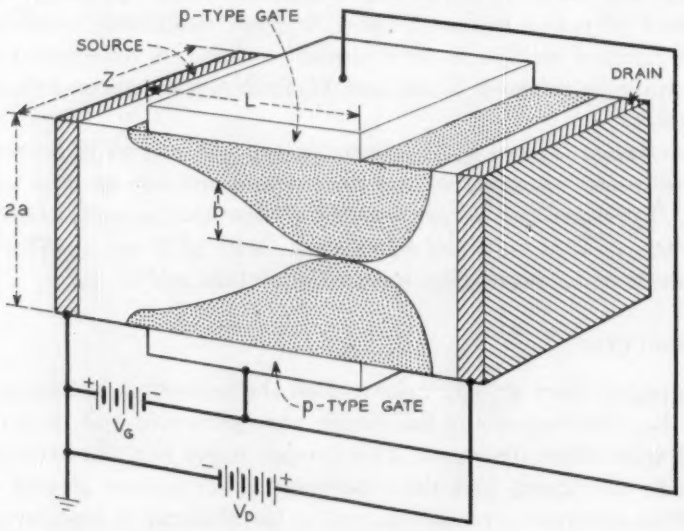


Fig. 1 — Schematic diagram of the field-effect transistor.

is the source of these electrons and the right hand contact drains them out of the material. The contacts are therefore referred to as "the source" and "the drain" and are marked accordingly in the figure. Because the n-type material has resistance, the flow of current will produce an IR drop, the potential becoming more positive toward the drain; but, since the gates are connected to the source, the same potential will appear across the p-n junctions and will bias them in the reverse direction. Space-charge regions will penetrate into the n-type material, becoming wider towards the drain since higher bias results in thicker space-charge regions. In consequence the current is confined to flow in a wedge-shaped region which is called "the channel". As the drain voltage is increased, the channel becomes narrower and the source to drain resistance higher, until at a voltage W_0 the condition is reached in which the space-charge regions from opposite gates meet. This is the condition represented in Fig. 1 where the channel is shown white and the space-charge regions are dotted. This is called "the pinch-off" condition. Beyond pinch-off the current is essentially saturated at a value I_{D0}, further increase of drain voltage resulting in only small increases in current, because most of the increased voltage appears across the space-charge region near the drain. The drain characteristic is thus of the form shown in the upper curve of Fig. 2.

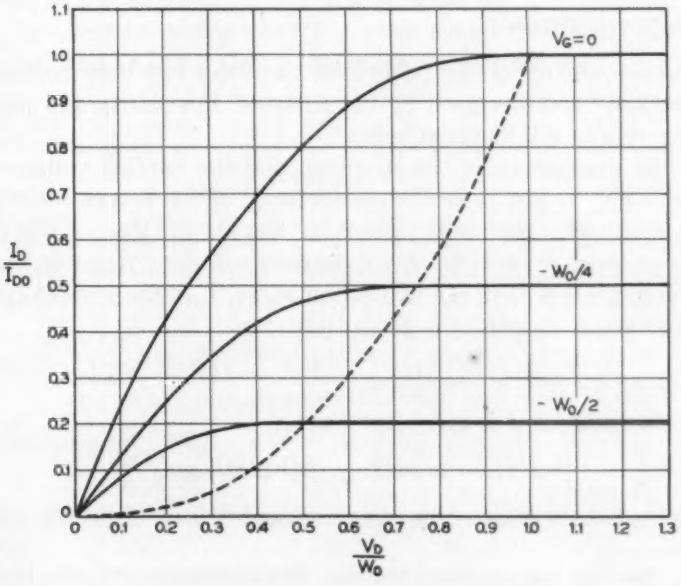


Fig. 2 — Theoretical drain characteristics for the constant-mobility case.

Now suppose that the gates are biased negatively with respect to the source with a voltage V_G. Under these conditions the magnitude of the IR drop necessary to produce pinch-off will be smaller since part of this voltage is already supplied by the bias and current saturation will accordingly occur at lower values of drain voltage and current. The drain characteristics with gate voltage as a parameter are therefore of the form shown in Fig. 2. These characteristics are similar to those of a pentode tube, and the field-effect transistor can be similarly used to obtain amplification.

The operation of the field-effect transistor depends essentially on the presence of just one type of carrier, the majority carrier. For this reason it may be called a "unipolar" transistor. The junction transistor, however, depends for its operation on both types of carrier and is hence a bipolar transistor. In the field-effect, current is carried by the majority carrier drifting in an electric field, while for a junction transistor the current is carried by the minority carrier diffusing in an essentially field-free region. Since drift velocities can be very much higher than diffusion velocities, the transit time for similar dimensions can be shorter in the field-effect than in the junction type. Hence we would expect that with similar dimensions the field-effect device would be capable of operating at higher frequencies. This conclusion is borne out by the theory.

3. THE IDEAL THEORY

The basic theory of the field-effect transistor has been presented by W. Shockley[1] and discussed by the authors.[2] For clarity and completeness the results will be given here.

Let the dimensions of the specimen and the applied voltages be as shown in Fig. 1. Let σ_0 be the conductivity of the n-type material and ρ_0 the donor charge-density. Shockley has shown that, if the wedge-shaped channel referred to above narrows sufficiently slowly, then, for drain voltages less than the saturation value, V_D, the current (per unit length in the Z direction) is given by*

$$I_D = (1/L)\,[J(V_D - V_G) - J(V_s - V_G)] \tag{3.1}$$

where the function J is given by

$$J(x) = 2\sigma_0 a x[1 - (\tfrac{2}{3})\,(2xK/\rho_0 a^2)^{1/2}] \tag{3.2}$$

in which x has the dimensions of voltage and K is the dielectric constant.

* W. Shockley has discussed the case of a p-type channel. For the n-type channel appropriate changes of sign have been made.

For germanium

$$K = 1.42 \times 10^{-12} \text{ farads/cm}$$

Above the saturation voltage, V_D, the current is substantially constant

It will be convenient to introduce certain natural parameters defined as follows

$$\begin{aligned} W_0 &= \rho_0 a^2/2K \\ E_0 &= \rho_0 a/K = 2W_0/a \\ g_0 &= 2\rho_0\mu_0 a = 2\sigma_0 a \\ I_0 &= g_0 E_0 = 2\rho_0^2\mu_0^2/K \\ \tau_0 &= a/\mu_0 E_0 = K/\mu_0\rho_0 = K/\sigma_0 \end{aligned}$$

Note that W_0 is the bias voltage required between gate and channel to produce a space-charge region of thickness a. Hence, with the gates shorted to the source, W_0 is the saturation value of V_D. g_0 is the conductivity of the channel for zero gate and drain voltages. For small applied voltages the resistance, R_0, of the channel is

$$R_0 = L/2a\sigma_0$$

In Fig. 2 we have shown in terms of these reduced parameters the drain voltage-current characteristics as obtained from (3.1) and (3.2). From (3.1) and (3.2) we obtain the transconductance g_m:

$$\begin{aligned} g_m = \left.\frac{\partial I_D}{\partial V_G}\right|_{V_D=\text{const}} \\ = (2\sigma_0 a/LW_0^{1/2})[(V_D - V_G)^{1/2} - (V_s - V_G)^{1/2}] \end{aligned} \tag{3.3}$$

Now if in particular we take $V_s = 0$ and $(V_D - V_G) = W_0$ corresponding to grounded source and operation in the pinch-off region, (3.3) reduces to

$$\begin{aligned} g_{mG} &= (2\sigma_0 a/L)\,[1 - (-V_G/W_0)^{1/2}] \\ &= g_{m0}[1 - (-V_G/W_0)^{1/2}] \end{aligned} \tag{3.4}$$

where we have introduced

$$g_{m0} = \frac{2\sigma_0}{L} \tag{3.5}$$

the maximum transconductance.

The saturation drain current I_{DG} which flows for any gate voltage

can be obtained from (3.1) and (3.2) by setting $V_s = 0$, $(V_D - V_G) = W_0$. Doing this yields

$$I_{DG} = (g_{m0}W_0/3)\,[1 + (V_G/W_0)\,(3 - 2\sqrt{-V_G/W_0})] \tag{3.6}$$

We see from (3.6) that the current is completely cut off for a gate voltage of $V_G = -W_0$. Maximum current is

$$I_{D0} = g_{m0}W_0/3 \tag{3.7}$$

obtained for zero bias on the gate.

The frequency response of the device can be estimated by the following simple argument. In order to change the gate voltage, the capacity of its p-n junctions must be charged through the resistance of the channel. This process has an associated time constant which limits the frequency response. Let us assume a wedge-shaped channel, completely pinched off at the drain end and completely open at the source end (that is $V_G = V_s = 0$). The capacity for unit length in the Z direction is approximately

$$C \simeq 4KL/a \tag{3.8}$$

The factor 4 arises because the average width of the space-charge region is approximately $a/2$ and because there are two such regions, one on either side. This capacity on the average charges through half the resistance of the channel, i.e.,

$$R = L/2a\sigma_0 \tag{3.9}$$

We would accordingly expect a limiting frequency f given by

$$f = \frac{1}{2\pi RC} = \frac{1}{2\pi}\left(\frac{a^2\sigma_0}{2L^2K}\right) \tag{3.10}$$

Another way of looking at the frequency response is to consider the transit time of a carrier along the channel. In Appendix 1 it is shown that the transit time τ is given by

$$\tau = (3/2)\,(L^2/\mu_0 W_0) \tag{3.11}$$

Substituting for W_0, we obtain

$$\tau = \frac{3KL^2}{\sigma_0^2} \tag{3.12}$$

This transit time differs from RC in equation (3.10) by about a factor of $3/2$. Thus there is essential agreement between the frequency responses as estimated from RC and from transit time.

The above theory has been summarized in the form of nomographs

which are given in Figs. 3 and 4. In Fig. 3 we have given a nomograph for the calculation of the pinch-off voltage W_0. If a straight line is drawn between the value of N on the left scale and the value of a on the right scale, it intersects the center scales at W_0 and at E, the maximum value of the electric field in the space charge region. Fig. 4 shows simultaneously the field-effect parameters, a/L, σ_0, g_{m0}, R_0, f and C. A straight line drawn across the chart intersects the various scales in a set of values which are consistent with a given design. The one remaining parameter of interest, the current, is easily obtained from (3.7).

4. MODIFICATIONS OF THE IDEAL THEORY

The theory presented in Section 3 deals with an ideal structure. In practice we shall find it necessary to modify the theory somewhat to

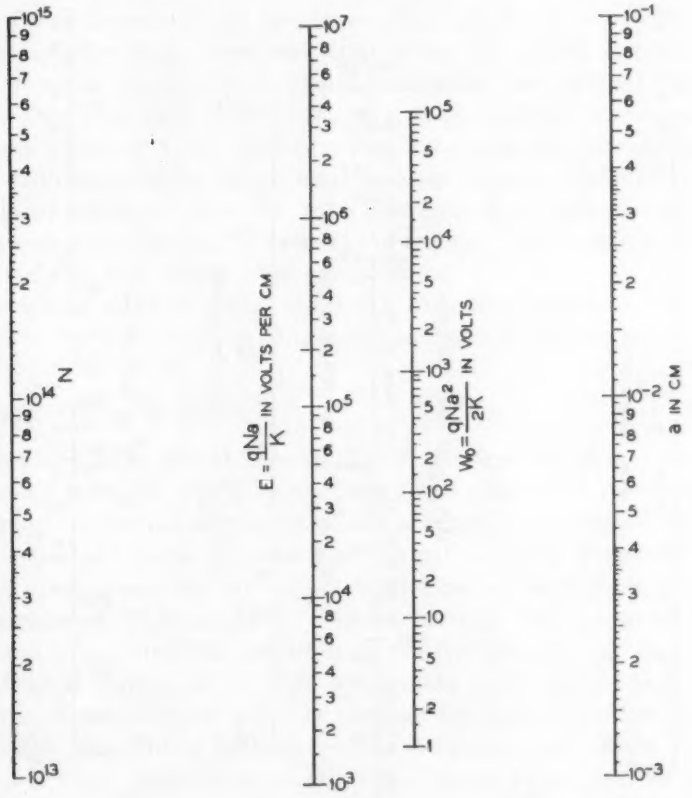


Fig. 3 — Nomograph giving the value of pinch-off voltage as a function of channel resistivity and thickness. (Constant-mobility case).

take account of special experimental conditions. In particular we shall discuss the following effects:

A. Series resistance of semiconducting paths at the source and/or drain contacts.
B. Negative resistance effects due to hole current flow into the gate.
C. Temperature effects.
D. Effects of high electric field on mobility.

A. Series Resistance

In Section 3 we have considered that the source and drain connect directly onto the channel between the gates. It is necessary in the fabrication of these units, however, to allow a small bridge of semiconductor between the actual contact and the gate. This means that a series re-

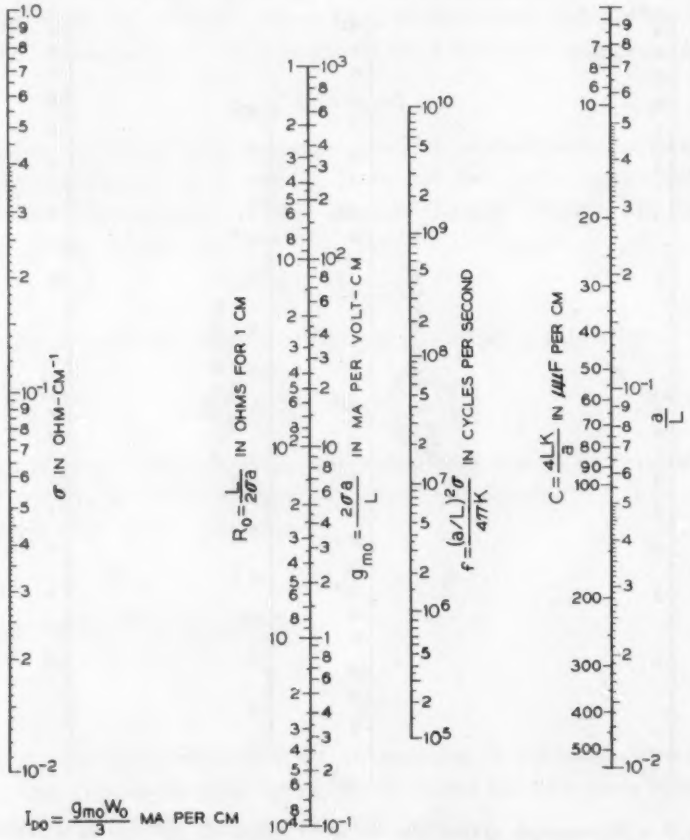


Fig. 4 — Design nomograph for constant mobility case.

sistance appears between the electrodes to which voltages are applied and the working part of the structure. It is possible to take account of these resistances by simple circuit theory.

For the calculation of g_m one must recognize two differences. Firstly, the voltages which must be inserted into (3.3) to obtain the transconductance of the working part of the structure must be changed as follows:

$$V_G \to V_G\,, \qquad V_s \to -IR_s\,, \qquad V_D \to (V_D - IR_D) \tag{4.1}$$

where R_s and R_D are the values of the source and drain resistances. Secondly, the degeneration of the source resistance must be taken into account. If the apparent transconductance is denoted by g_m' then

$$g_m' = g_m/(1 + R_s g_m) \tag{4.2}$$

which takes account of the fact that a fraction of any increase in gate voltage appears as an increased drop across the source resistance. In (4.2), of course, g_m must be calculated considering the modifications (4.1).

It is clear from (4.2) that if $g_m R_s \gg 1$, then the transconductance g_m' is simply given by $1/R_s$. If then we are to obtain the potentially high transconductance of the device, we must keep the source resistance small. The drain resistance does not have as serious consequences, the chief disadvantages being: (a) the necessity of having a higher supply voltage, and (b) $I_D^2 R_D$ heat which must be dissipated.

The source resistance also alters the frequency response. One must add R_s to the R of (3.10) in obtaining the limiting frequency.

B. Negative Gate Resistance

It is possible, in units where an appreciable fraction of the current is carried by holes, as would be the case for a channel of resistivity approaching the intrinsic value, to obtain a negative gate-resistance. This effect arises as follows: Suppose that the gate is made more positive so that it opens somewhat and allows additional electron current to flow in the channel. This additional electron current is accompanied by an increased hole current on the drain side of the gate, (within the electrically neutral region). If the unit is operating in the pinched-off region, however, these holes will not flow through the channel but will instead flow to the gate. This is the case because the electric field in the pinched-off space-charge region near the drain is directed away from the axis of the channel and is thus focussing for electrons, while tending to pull holes into the gate. From this argument we see that a positive change of gate voltage results in an increased flow of holes to the gate and thus

in a negative change in the current flowing into the gate. Hence the gate terminal presents a negative resistance.

A quantitative estimate of the negative resistance can be obtained as follows: let the fraction of the drain current which is carried by holes be f. Let the gate potential increase by V volts. If the transconductance is g_m, the drain current, I_D, will increase by $g_m V$. Then the hole current will increase by $fg_m V$. All this hole current appears as a current flowing out of the gate. Thus the change in gate current is $-fg_m V$ and the gate conductance $G_G = -fg_m$.

The holes referred to in the argument above may arise from three sources: (a) they may be thermally generated in the body of the semiconductor; (b) they may be generated at low lifetime areas on the surface; or (c) they may be injected at the contacts. It might be desirable to make a device in which the drain contact was intentionally made hole-injecting (say by the use of a p-n junction) and in which the negative resistance effects would thus be greater. However, for most applications the negative resistance is not desirable, and it is necessary to minimize the density of these holes. Source (a) can be decreased by using lower resistivity material, and source (b) by high lifetime treatment of the surface. The injection (c) of holes at the drain contact can be prevented by making the drain contact from strongly n-type material (designated n^+). Such an n-n^+ junction will not act as a source of holes.

C. Temperature Effects

As can be seen from the nomographs, the unipolar field effect transistor is a relatively high power device. Furthermore, most of the power dissipation takes place in the space-charge region near the drain. Therefore the removal of heat can be a major problem. The chief effect of heating is to increase the gate saturation current by increasing the number of thermally generated hole-electron pairs. It is true that the mobility also varies as $T^{-3/2}$ but this variation is fairly slow and serves only to vary σ_0 which enters the theory linearly.

D. Effects of Non-linear Mobility at High Fields

In germanium there is a maximum field E_c beyond which Ohm's law fails.[3] Beyond E_c the mobility decreases as the half power of the electric field, and hence the effective conductivity of the material also changes. As will be shown in Part 2, electric fields greater than E_c may well be obtained in practical designs of field-effect transistors. It is therefore necessary to consider the effects of square-root mobility on the tran-

sistor characteristics. Since the required analysis is somewhat lengthy, and the conclusions that can be drawn from it are of great interest, this analysis will be given in a separate section.

5. SQUARE-ROOT MOBILITY THEORY

The treatment which follows is similar in outline to that of W. Shockley for the constant-mobility case[1] and familiarity with the general features of that theory will be assumed. The fundamental assumption made is that the channel narrows sufficiently slowly so that locally the voltage W across the space-charge region between the channel and the gate is given by the following solution to the one-dimensional Poisson equation

$$W = qN[a - b(x)]^2/2K \tag{5.1}$$

where q is the electronic charge, N is the excess donor densitiy* in the channel material, 2a is the thickness of the material between the gates and 2b(x) is the thickness of the narrowed channel at a distance x from the source towards the drain. This situation is shown in Fig. 1.

We shall consider the case where the pinch-off voltage is so high that the field in the channel is greater than the critical field over the major fraction of its length, but not so high that the carriers reach limiting velocity.[3] A more accurate calculation would take into account the fact that near the source the field is small and would join two solutions for the ohmic and non-ohmic parts. We shall assume that in the range for which $E > E_c$, the mobility is proportional to $(E)^{-1/2}$. This law is in agreement with the data of E. J. Ryder.[3] In particular we write for the conductivity in the channel

$$\sigma_c = \mu_0 nq(E_c/E)^{1/2} = \sigma_0(E_c/E)^{1/2} \tag{5.2}$$

where μ_0 is the low field mobility, n the electron density in the channel, E_c the critical field, and σ_0 the low field conductivity.

We shall consider a unit 1 cm wide, $2a$ cm thick between gates, and with a channel length of L cm. The conductance g of the channel at any point x is proportional to the thickness of the channel there and is given by

$$g = 2b(x)\sigma_c = 2b(x)\sigma_0(E_c/E)^{1/2} \tag{5.3}$$

Making use of (2.1) we may write this in terms of W

$$g = 2\sigma_0 a[1 - (W/W_0)^{1/2}]\,(E_c/E)^{1/2} \tag{5.4}$$

* The theory is worked out here for a transistor with n-type channel and p-type gate. This is the opposite polarity to that treated by Shockley in Reference 1, but is appropriate to the specimens measured.

where $W_0 = qNa^2/2K$ has been introduced. The gradual approximation implies that $E = dW/dx$, and we may therefore write for the square of the current

$$I^2 = g^2(dW/dx)^2 = 4\sigma_0^2 a^2 E_c[1 - (W/W_0)^{1/2}]^2(dW/dx) \tag{5.5}$$

Upon integration from source to drain we obtain

$$I^2 L = g_0^2 E_c \int_{W_S}^{W_D} [1 - (W/W_0)^{1/2}]^2 \, dW \tag{5.6}$$

where we have introduced the symbol $g_0 = 2\sigma_0 a$. We shall find it convenient to introduce the integral

$$J(W) = \int_0^W [1 - (y/W_0)^{1/2}]^2 \, dy \tag{5.7}$$

in terms of which the current may be written

$$I = g_0(E_c/L)^{1/2}[J(W_D) - J(W_S)]^{1/2} \tag{5.8}$$

It is easy to evaluate $J(W)$ by the change of variable

$$u^2 = [1 - (y/W_0)^{1/2}]$$

and when this is done we find

$$J(W) = \frac{W_0}{6}(3[1 - (W/W_0)^{1/2}]^4 - 4[1 - (W/W_0)^{1/2}]^3 + 1) \tag{5.9}$$

We are now in a position to write the current from (5.8) and (5.9). We shall take

$$W_S = V_S - V_G\,; \qquad W_D = V_D - V_G \tag{5.10}$$

where V_S, V_D, and V_G are the potentials applied to the source, drain, and gate respectively. We accordingly obtain

$$\begin{aligned} I &= I_c\{3[1 - (V_D - V_G)^{1/2}/W_0^{1/2}]^4 - 4[1 - (V_D - V_G)^{1/2}/W_0^{1/2}]^3 \\ &= 3[1 - (V_S - V_G)^{1/2}/W_0^{1/2}]^4 + 4[1 - (V_S - V_G)^{1/2}/W_0^{1/2}]^3\}^{1/2} \end{aligned} \tag{5.11}$$

where we have introduced the symbol

$$I_c = g_0(W_0 E_c/6L)^{1/2} \tag{5.12}$$

This is the value of the current at pinch-off for zero gate bias. The analogous expression for the constant mobility case was[2]

$$I_{D0} = g_{m0}W_0/3 = 2W_0\sigma_0 a/3L = W_0 g_0 3L \tag{5.13}$$

Therefore, all other things being equal, the effect of non-constant mo-

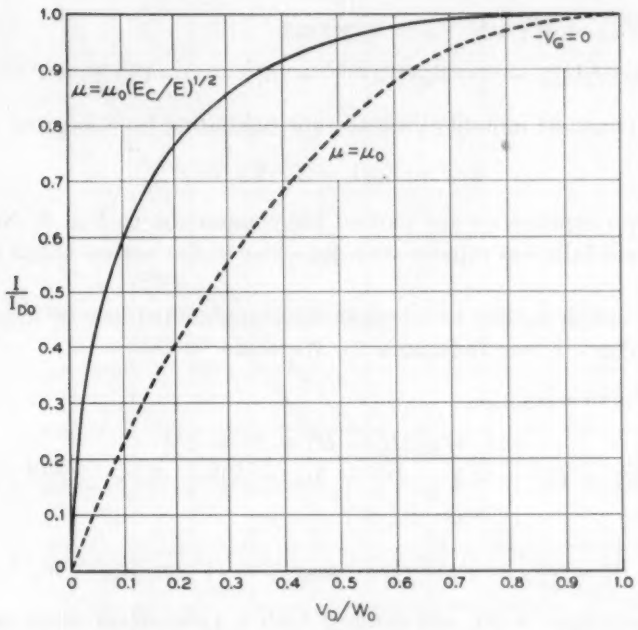


Fig. 5 — Comparison of the drain characteristics for the constant and non-constant mobility cases.

bility is to reduce the maximum current in the ratio

$$I_c/I_{D0} = (9E_cL/6W_0)^{1/2} = 1.21\ (E_cL/W_0)^{1/2} \tag{5.14}$$

It should be pointed out that this analysis is valid only for $W_0/L \gg E_c$, but within this range of validity, equation (5.14) shows that considerable reductions of current may take place.

In addition to changes in magnitude of the pinch-off current, there will be significant alterations in the shape of the I_D versus V_D characteristic. This characteristic has been calculated from equation (5.11) for the case $V_S = V_G = (V_S - V_G) = 0$ and is plotted in Fig. 5. Also shown on the same figure, for comparison purposes, is the I_D versus V_D curve for the constant mobility case. It can be seen that the non-constant mobility curve has a steeper initial slope and pinches off more gradually. The effect of this is to make the unit *appear* to have a lower pinch-off voltage. At low drain voltages, on the other hand, the unit is operating in the Ohm's law range and the observed initial resistance will agree with the slope of the curve for the constant mobility case.

It is also of interest to know the variation of the current beyond pinch-off with gate bias. This function can be obtained from (5.11) by putting

$(V_D - V_G) = W_0$ and $V_S = 0$, giving

$$I_{cG} = I_c(4[1 - (-V_G/W_0)^{1/2}]^3 - 3[1 - (-V_G/W_0)^{1/2}]^4)^{1/2} \quad (5.15)$$

For the constant mobility case the corresponding function was

$$I_{DG} = I_{D0}[1 - (-V_G/W_0)^{1/2}] \quad (5.16)$$

These two expressions are plotted for comparison in Fig. 6. Note that the current falls less rapidly with gate bias in the non-constant mobility case.

It is a simple matter to calculate the transconductance by differentiating the current (see Reference 2). We write

$$g_m = \partial I_D/\partial V_D \,|_{V_D = \text{const}}$$

$$= \frac{(3I_c/W_0)[2D - D^2 + S^2 - 2S]}{[3(1 - D)^4 - 4(1 - D)^3 - 3(1 - S)^4 + 4(1 - S)^3]^{1/2}} \quad (5.17)$$

where

$$D = (V_0 - V_G)^{1/2}/W_0^{1/2} \text{ and } S = (V_S - V_G)^{1/2}/W_0^{1/2}$$

In particular, if we are dealing with a pinched-off drain and zero gate bias, that is $(V_D - V_G) = W_0$ and $(V_S - V_G) = 0$, then both of the brackets in (5.17) reduce to unity and we obtain

$$g_{mc} = 3I_c/W_0 \quad (5.18)$$

The analogous expression for the constant mobility case was

$$g_{m0} = 3I_{D0}/W_0 \quad (5.19)$$

Therefore we see that the effect of square-root mobility is to reduce the transconductance in the same ratio as the current, that is,

$$g_{mc}/g_{m0} = I_c/I_{D0} = 1.21\ (E_cL/W_0)^{1/2} \quad (5.20)$$

It is also of interest to consider the variation of transconductance with gate bias. For this purpose in (5.17) we set $(V_D - V_G) = W_0$ and $V_S = 0$, giving

$$g_{mG} = g_{mc}[1 - (-V_G/W_0)^{1/2}]^{1/2}[1 + 3(-V_G/W_0)^{1/2}]^{-1/2} \quad (5.21)$$

The analogous expression for the constant mobility case was

$$g_{mG} = g_{m0}[1 - (-V_G/W_0)^{1/2}] \quad (5.22)$$

In Fig. 7 we have plotted curves of g_{mG}/g_{mc} and g_{mG}/g_{m0} for the non-constant and constant mobility cases versus V_G/W_0. It can be seen that

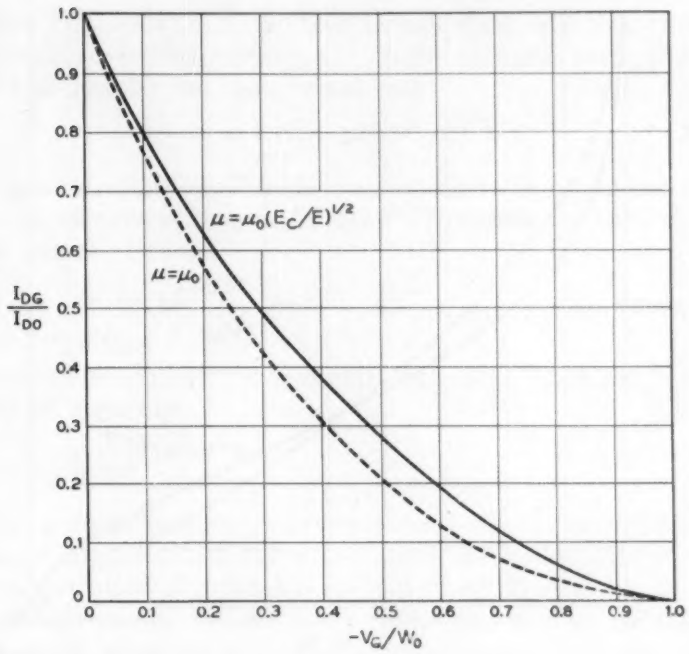


Fig. 6 — Comparison of gate characteristic for constant and non-constant mobility cases.

the effect of the non-constant mobility is to introduce an initially more rapid fall in transconductance with gate bias, followed by a relatively slower fall. The initial rapid fall of transconductance suggests that the gate bias must be held very close to zero if the maximum transconductance is to be observed. In practice, it may not be possible to achieve this because of voltage drop in the resistance of the source or gate leads. In any case, this effect will constitute a rather serious limitation on the circuit behavior of the device. In order to maintain large transconductance it is necessary to employ small bias and small signal-swing.

An approximation to the frequency response for the constant mobility case was obtained by Shockley[1] from the assumption that the capacity between the channel and gate must be charged through the resistance of the channel. Approximating the channel shape at pinch-off by a triangular wedge, he obtained

$$f = \sigma_0(a/L)^2/4\pi K \tag{5.23}$$

In the non-constant mobility case a similar sort of approximation could be made. The channel will also be roughly wedge shaped (see

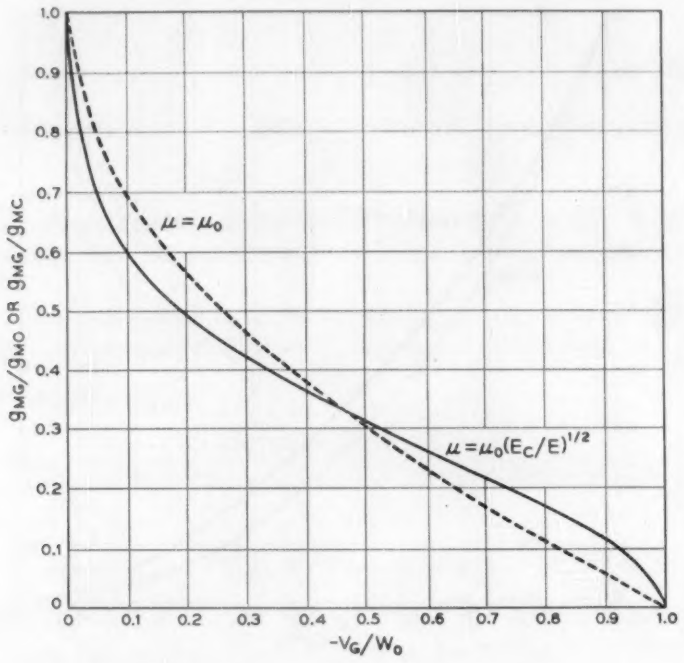


Fig. 7 — Dependence of g_m on V_g. The dotted line refers to the constant-mobility case and the solid line to the non-constant case.

Appendix 1). The conductivity however will vary along the channel as the electric field varies. We may within the accuracy of (5.23) use an average conductivity corresponding to the average field, $E_a = W_0/L$, in the channel. Making use of (5.2) we find that the decrease in frequency response caused by non-constant mobility can be estimated by

$$f_c/f = \sigma_c/\sigma_0 \approx (E_c/E_a)^{1/2} = (E_cL/W_0)^{1/2} \tag{5.24}$$

The quantity $(E_cL_i/W_0)^{1/2}$ is seen to be a sort of figure of comparison for the non-constant mobility case. Except for numerical factors near unity, the current, the transconductance, and the frequency response are all reduced in about this same ratio

$$I_c/I_{D0} \approx g_{mc}/g_{m0} \approx f_c/f \approx (E_cL/W_0)^{1/2} \tag{5.25}$$

Another way of looking at the frequency response is to consider the transit time across the channel. In Appendix 1 it is shown that for the constant mobility case

$$\tau = (3/2)\ (L^2/\mu_0 W_0) \tag{5.26}$$

Note that $L^2/\mu_0 W_0$ is the transit time across a distance L that we would expect from the average velocity $\mu_0 W_0/L$. On the other hand, as shown in the Appendix, for the non-constant case

$$\tau_c = 1.47\ L/\mu_0(E_c W_0/L)^{1/2} \tag{5.27}$$

Here again $L/\mu_0(E_c W_0/L)^{1/2}$ is the transit time to be expected across a distance L with the velocity $\mu_0(E_c W_0/L)^{1/2}$. Note that the ratio of (5.26) to (5.27) gives

$$\frac{f_c}{f} = \frac{\tau}{\tau_c} = \frac{1.5}{1.47}\frac{L^2}{\mu_0 W_0}\cdot\frac{\mu_0(E_c W_0/L)^{1/2}}{L} = 1.02(E_c L/W_0)^{1/2} \tag{5.28}$$

This expression is in essential agreement with (5.25) which was obtained from the *RC* argument.

6. DESIGN THEORY

In the previous sections we have presented the theory of field-effect transistors operating both below and above the critical field E_c. The question now arises as to what is the optimum field at which the transistor should operate. We will first consider this problem for the constant mobility range and show that both frequency response and transconductance increase with electric field. We shall then consider the square-root theory and show that, although increasing field above E_c does result in an increase in frequency response and transconductance, the corresponding increases in power dissipated are so large as to make such a design unattractive. These considerations lead to the conclusion that the optimum field for transistor operation is just the critical field, and we will present design nomographs for this condition.

A. Constant Mobility

For the gradual approximation, that is $(a/L) \ll 1$, the following equations apply

$$f = (a/L)^2\sigma_0/4\pi K \tag{6.1}$$

$$W_0 = qNa^2/2K \tag{6.2}$$

$$g_{m0} = 2\sigma_0(a/L) \tag{6.3}$$

where the symbols are as previously defined. It is seen from (6.1) and (6.3) that both the frequency response and transconductance are proportional to conductivity. Thus the designs of most interest are those with high conductivity material. The solution of the three equations

would be much simplified if it could be assumed that σ_0 is proportional to N. This is true if the density of the minority carrier is small compared to that of the majority carrier and is therefore true for material of high conductivity, corresponding to the cases of greatest interest. If we assume that

$$\sigma_0 = q\mu_0 N \tag{6.4}$$

the error introduced will be about 10 per cent for $\sigma_0 = 0.05$ ohm^{-1}-cm^{-1} and less than 1 per cent for $\sigma_0 > 0.7$ ohm^{-1}-cm^{-1}. It is therefore reasonable to apply (6.4) with the restriction that

$$\sigma_0 > 0.05 \text{ ohm}^{-1}\text{-cm}^{-1}$$

The additional restriction that the field in the channel must be less than the critical field is difficult to apply exactly. The field in the channel increases from the source to a maximum at the drain. The dependence of field upon distance down the channel has been derived by Shockley, but it is a complicated expression. A simple approximation to the limiting condition would be to stipulate that the average field in the channel, defined as $E_a = W_0/L$, be less than E_c. In the limiting case, $W_0/L = E_c$, part of the channel near the drain would be above the critical field while the source end would not. This represents approximately the desired limiting condition.

We shall now impose onto (6.1) to (6.4) the following three restrictions:

(a) Gradual approximation holds,

$$a/L \ll 1$$

(b) Conductivity is proportional to majority carrier density,

$$\sigma_0 > .05 \text{ (i.e., } \rho < 20 \text{ ohm-cm)}$$

(c) Field is less than critical,

$$W_0/L = E_a \leqq E_c$$

Firstly, consider the frequency response. Eliminating σ_0 and N from (6.1), (6.2) and (6.3) gives

$$f = (a/L)^2 \mu_0 W_0/2\pi a^2 = \mu_0 W_0/2\pi L^2$$

and substituting for W_0 from restriction iii yields

$$f = \mu_0 E_a/2\pi L \tag{6.5}*$$

* Note that this equation shows that f is proportional to $\mu_0 E_a$, the mean velocity of the electrons, divided by L, the length of the channel. Since E is always positive and increases smoothly from source to drain, the ratio $\mu E_0/L$, and hence f, will be approximately proportional to the inverse of the transit time. This result is rigorously shown in Appendix 1.

This equation shows that the frequency response is proportional to the average field and that therefore for a given length of channel the maximum frequency response will be obtained with the maximum possible value of E_a, that is, $E_a = E_c$. Another interesting relationship can be deduced from (6.5) by substituting for L from restriction (c) giving

$$f = \mu_0 E_a^2 / 2\pi W_0 \tag{6.6}$$

Thus for fixed value of E_a high values of frequency response will correspond to low values of pinch-off voltage.

Now consider the transconductance. Eliminating σ_0 and N from (6.2), (6.3) and (6.6) gives

$$g_{m0} = 4\mu_0 K W_0 / aL$$

and substituting from restriction iii for W_0/L, we obtain

$$g_{m0} = 4\mu_0 K E_a / a \tag{6.7}$$

Thus for a given thickness of channel the maximum transconductance will be obtained with the maximum possible value of E_a, that is $E_a = E_c$.

We therefore conclude that on the constant-mobility theory the best choice of average field for optimum frequency response and transconductance is the critical value, E_c.

B. Square-Root Range

Having shown that on the constant mobility theory the optimum electric field is E_c, we must next determine what, if anything, is to be gained by operating above this field; that is, in the square root mobility range. It is clear from (5.24) that, if units are operated at higher and higher W_0, the gain in frequency response is not as great as would be predicted by the constant mobility theory. In fact for a given length L, the frequency response can never exceed

$$f_t = 1/\tau_t = v_t/L \tag{6.8}$$

where v_t is the limiting drift velocity of "hot" electrons.[3] It should be pointed out however, that constant electron drift velocity is incommensurate with the usual gradual-approximation field-effect equations. This condition can obtain only at the pinched-off end of the channel, i.e., the "expop" region, for within the electrically neutral channel continuity of current and constant electron drift-velocity combine to require constant b. Fulfillment of this condition is impossible between equipotential gates since the change in W with x required to give the electric field would also require a change in b by (5.1).

As we shall subsequently show, it is not expedient to attempt to obtain high frequencies by the brute-force method of increasing W_0/L because the power goes up as W_0^2, while square-root mobility effects cause diminishing returns to set in as far as increases in frequency response are concerned. A simple way to see how much improvement in frequency response can be obtained by pushing into the nonlinear mobility range is to impose a maximum power dissipation condition. It has been found possible experimentally to dissipate by the use of cooling fins some 400 watts per cm length of gate. We will accordingly take $P_c = I_c W_0/L = 400$ as a limiting power-handling capacity. We write the square-root case equations in the following form:

$$g_{mc} = 6^{1/2}\sigma_0(L/a)^{-1}L^{1/2}W_0^{-1/2}E_c^{1/2} \tag{6.9}$$

$$f_c = \sigma_0(L/a)^{-2}L^{1/2}W_0^{-1/2}E_c^{1/2}(1/4\pi K) \tag{6.10}$$

$$P_c = I_c W_0/L = g_{mc}W_0^2/3L \tag{6.11}$$

$$W_0 = \sigma_0(L/a)^{-2}L^2\left(\frac{1}{2\mu_0 K}\right) \tag{6.12}$$

Using (6.12) and (6.9) to write P_c in terms of ρ, (L/a), and L we obtain

$$P_c = \sigma_0^{5/2}(L/a)^{-4}L^{5/2}E_c^{1/2}\left(\frac{6^{1/2}}{3}\right)\left(\frac{1}{2\mu_0 K}\right)^{3/2} \tag{6.13}$$

We now use (6.12) and (6.13) to eliminate W_0 and σ_0 from (6.10) and finally obtain

$$f_c = (1.25P_cE_c^2\mu_0^4/128\pi^5K)^{1/5}(L/a)^{-1/5}L^{-1} \tag{6.14}$$

If, in agreement with experimental results, we take

$$P_c = 400 \text{ watts/cm}$$

$$E_c = 1000 \text{ volt/cm}$$

$$\mu_0 = 3600 \text{ cm}^2/\text{volt-sec.}$$

and

$$K = 1.4 \times 10^{-12} \text{ farad/cm}$$

(6.14) becomes

$$f_c = 1.05 \times 10^6(L/a)^{-1/5}L^{-1} \tag{6.15}$$

It should be noted that (6.15) is relatively insensitive to choices of (L/a).

For any (L/a) from 1 to say 5, (6.15) can be approximated by

$$f_c \approx 1.05 \times 10^6 L^{-1} \tag{6.16}$$

On the other hand (6.8) gives

$$f_t = 6 \times 10^6 L^{-1} \tag{6.17}$$

when Ryder's experimental value for v_t is inserted. It is clear, therefore, that the power handling capacity of a unit of given size prevents the realization of the ultimate frequency response and causes a reduction of upper frequency limit by a factor of about 6. From (6.14) we see that f_c depends on $P_c^{1/5}$ and that it is necessary to increase the power handling capacity by a factor of 32 in order to double the frequency response.

We have explored the design possibilities of units in which the critical field is never exceeded. It is shown that the limiting frequency under these restrictions is given by

$$f = E_c\mu/2\pi L = 5.7 \times 10^5 L^{-1} \tag{6.18}$$

A comparison of (6.16) and (6.18) shows that a gain of only a factor of 2 in frequency response can be obtained by operating in the non-constant mobility range. This factor is gained at a considerable cost in power dissipation, and it is therefore unprofitable to extend the design into this operating range.

Design Nomographs for $W_0/L = E_c$

Nomographs have been given, Fig. 3 and 4, representing the "field-effect" equations for the gradual approximation without any restrictions on field. It will be noted that, in order to obtain a unique set of solutions with these nomographs, three choices have to be made. For example, if σ_0 (and therefore N) and a are chosen, W_0 is determined in Fig. 3, but another choice, say (a/L), must be made to determine the values of f and g_{m0} in Fig. 4. If now we stipulate that $W_0/L = E_c$, we are left with only two choices and therefore the two nomographs can be combined in one. This new nomograph is shown in Fig. 8. In determining the scale factors $E_a = E_c$ was taken as 1,000 volts/cm. For n-type germanium at room temperature the critical field is actually 900 volts/cm but the figure of 1,000 volts/cm was used to give some simplification of the scale factors.

Of the three restrictions imposed on the analysis, restriction iii has been incorporated in the nomograph of Fig. 8. Restriction (a) states that (a/L) shall be small compared to unity. It is difficult to assess exactly

the error that will be introduced as (a/L) approaches unity but certainly the nomograph should not be used for values of (a/L) greater than ½. This region has been marked on Fig. 8. Restriction (b) states that the analysis will not be accurate for materials with $\sigma_0 < 0.05$ ohms^{-1}-cm^{-1} and this region has also been marked on Fig. 8.

A theoretical limitation which has not so far been considered is concerned with the minimum allowable pinch-off voltage. If the reverse gate bias is insufficient to saturate the junction, the gate impedance will be much lower than the value predicted from theory and will vary with gate bias. It is therefore necessary that the gate be saturated, i.e., that the reverse bias be large compared to kT/q, which, at room temperature, is approximately 0.025 volts. In practice it is found that a junction is sensibly saturated at 0.1 volts bias. However, if the gate is to be saturated

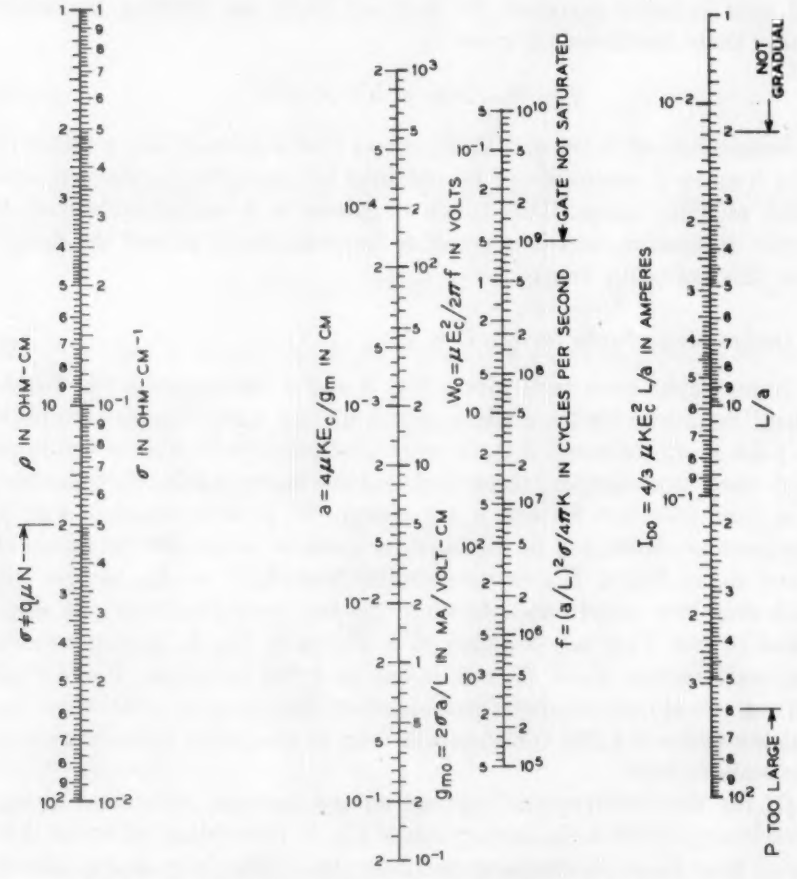


Fig. 8 — Design nomograph for average channel field equal to E_c.

over most of its length the pinch-off voltage, W_0, must be much greater than 0.1 volts. A reasonable lower limit to the value of W_0 would therefore be about 0.5 volt, and this limitation is marked on the nomograph. As a consequence of this choice, the upper limit of frequency response, as given by the nomograph, is 10^9 *c/s*.

A practical limitation in the design of field-effect transistors is the allowable power dissipation in the unit. This is determined by the power that can be removed without undue temperature rise. With use of a cooling fin on the gate, the power that can be removed from a unit of 1 cm width will be approximately proportional to the length, L, of the channel. Therefore, a reasonable design criterion would be the power dissipated per unit length of channel, that is, $P = W_0 I_{D0}/L$. This function reduces to

$$P = W_0 I_{D0}/L = (4/3)\mu_0 K E_c^3 (L/a)$$

Putting in numbers for μ_0, K and E_c gives

$$P = 6.7\ (L/a) \tag{6.19}$$

In the units that have been made and cooled by this method, the channel length was 0.005 inches. With the fin water-cooled a maximum power of 5 watts could be dissipated with only a small rise in temperature. If this case is taken to represent the best that can be done, then $P \doteq 400$ watts/cm and substituting in equation (6.19) gives a maximum value of (L/a) of 60. Therefore the limitation imposed by the allowable power dissipation restricts the designs to those having values of (L/a) less than 60. This region is marked on the nomograph.

Some Possible Designs of Field-Effect Transistors

In Table I are shown the properties of some transistors designed by means of the nomograph. Unit No. 1 is chosen to give the maximum theoretical frequency response, 1,000 mc/s, with the widest possible channel. It is therefore the most feasible design for the highest frequency response. However, the dimensions of this unit — channel length and distance between gates both being about ¼ mil — make fabrication difficult. If such a unit could be made, it would have some very desirable properties: frequency response of 1,000 mc/s, transconductance of 70 *ma/v* and ability to operate with a 12 ma current from a 0.5-volt supply.

Unit No. 2 is designed for maximum frequency response together with maximum transconductance. The dimensions of this unit are too small to be considered feasible with existing techniques. Unit No. 3, designed for maximum power dissipation could be made provided that

TABLE I — DESIGN PARAMETERS AND CHARACTERISTICS OF SOME FIELD EFFECT STRUCTURES

No.	Remarks	ρ	a	L	f	g_{m0}	W_0	I_{D0}	Power
		ohm-cm	*cm*	*cm*	*C/S*	*ma/v*	*Volts*	*ma*	*Watts*
1	Max. f, max. a	15	3×10^{-4}	6×10^{-4}	10^9	70	0.5	12	6×10^{-3}
2	Max. f, max. g_{m0}	0.07	2×10^{-5}	6×10^{-4}	10^9	10^3	0.5	180	9×10^{-2}
3	Max. power	20	1.1×10^{-3}	6.6×10^{-1}	8×10^5	1.7	600	400	240
4	$\rho = 20$, $W_0 = 30$	20	2.6×10^{-3}	2×10^{-2}	1.6×10^7	8	30	80	2.5

junctions to 20 ohm-cm germanium could be made with breakdown voltage greater than 600 volts. If it were possible to obtain body breakdown,[4] such junctions would stand approximately 800 volts reverse bias. In practice however, perhaps due to surface breakdown, it is difficult to exceed 100 volts.

Unit No. 4 represents about the best unit that has been made with the molten-metal process.

PART II. EXPERIMENTAL RESULTS

1.0 INTRODUCTION

In Part I we have presented the design theory of field-effect transistors; in this part we shall present experimental data which verify the design theory. All specimens to be described were made using n-type *Ge* for the channel material. The p-type gates were formed by the indium alloy-process.

In order to obtain higher frequency response the dimensions of the channel were made small. The exact choice of resistivity was determined by two considerations. It must be sufficiently low to avoid negative-resistance effects but not so low that the pinch-off voltage becomes excessive for reasonable channel thickness. It was found that 20 ohm-cm germanium was a good compromise. By the use of a confining jig during alloying, it was possible to make alloy gates as small as 5 mils long. With such dimensions and material, the power dissipation in the channel was of the order of a few watts and means had to be provided for removing heat from the specimen. This was done by attaching a cooling fin to the indium during the alloying process. As previously,[2] a grown n-n^+ junction was incorporated to provide a drain contact that would not inject

holes and a tin alloy contact was used for the source. Close source to gate spacing was achieved by alloying the source and gate contacts simultaneously in the jig. A schematic diagram of the unit is shown in Fig. 9 and a photograph of the completed structure in Fig. 10. An exploded view of the alloying jig used in making the unit is shown in Fig. 11.

In all, 9 successful units were made in this way. The salient properties of these are shown in Table II. Before discussing these properties we will describe the testing procedures used.

2. MEASUREMENT TECHNIQUES

2.1 *Measurement of Static Characteristics*

A preliminary examination of the characteristics was made using a pulse-operated *E-I* presentation unit. In this way it was possible to

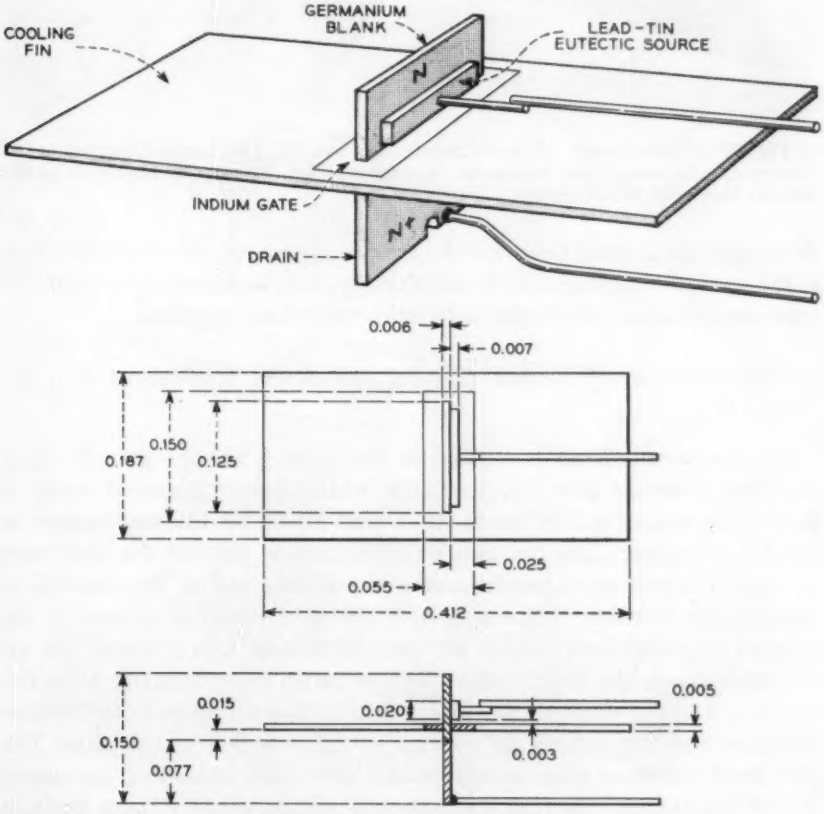


Fig. 9 — Schematic diagram of experimental transistor.

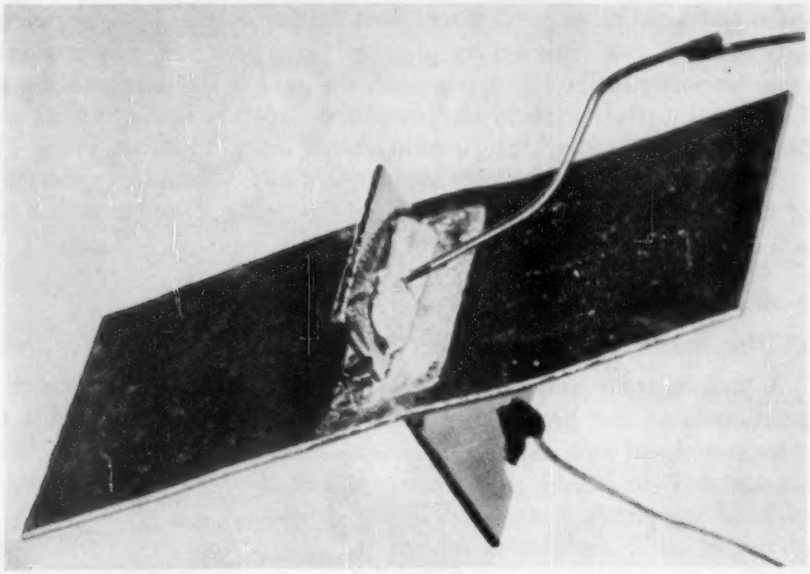


Fig. 10 — Photograph of experimental transistor. The largest component is the cooling fin which also serves as the gate contact. The upper lead goes to the source, the lower to the drain.

determine the general behavior of a specimen without risk of overheating. If the specimen appeared to be satisfactory, that is, showed pinch-off and transconductance, the dc characteristics were then measured.

2.2 *Determination of Source, Channel and Drain Resistances* R_s, R_0, *and* R_D

If a positive voltage is applied to the drain while the gate is open-circuited, then the gate will float at a potential approximately equal to that at the source end of the channel. The reason for this can be seen in the following way. The net gate current must be zero. If the gate were to become much more positive than the source end of the channel, a considerable fraction of the length of the gate would be biased in the forward direction and a large current would flow into the gate. If, on the other hand, the gate were to become much more negative than the source end of the channel, all of the gate junction would be in the reverse direction and the saturation current would flow out of the gate. The gate must therefore take up a potential very close to that of the source end of the channel. In fact the potential will be slightly more positive so that a forward current from a small portion of the gate at the source

end will neutralize the reverse saturation current of the remainder of the gate. Thus the open-circuit gate potential is closely equal to the potential drop in the source resistance R_S and, if the current I_S is measured, the value of R_S can be determined. In the experiment, the gate potential was measured with an electronic voltmeter having an input impedance of 15 $M\Omega$, so that the gate was essentially open-circuited.

To determine the value of R_D the source and drain connections were interchanged and the experiment repeated. R_0 was determined from the measured value of the sum of R_D, R_0 and R_D. The sum was determined from the ratio of V_D to I_D for values of V_D sufficiently small compared to W_0 so that there was negligible "pinching-off" in the channel.

2.3 *Frequency Response*

The cut-off frequency of many of these units was about 20 mc/s and in one case, as high as 50 mc/s. A thorough investigation of the frequency

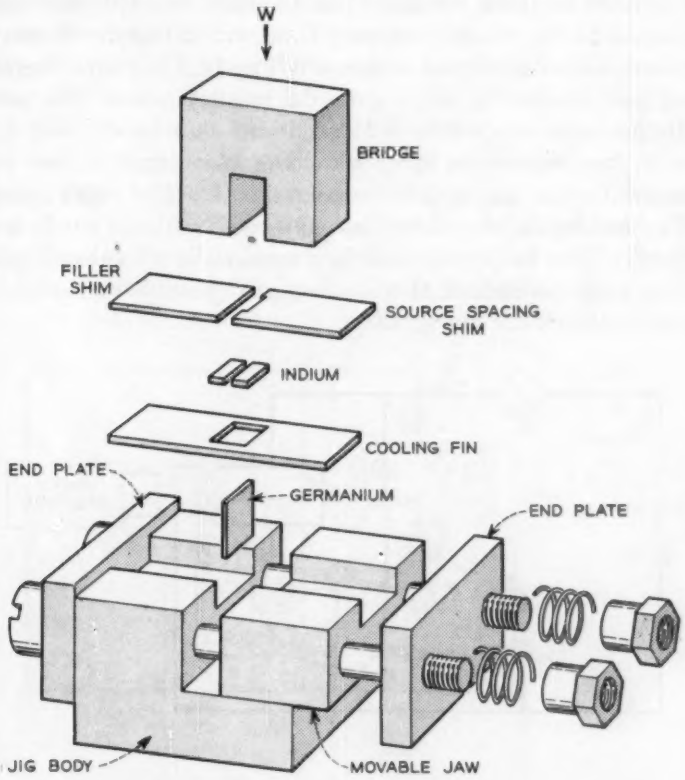


Fig. 11 — Exploded view of the jig used in fabrication.

TABLE II

Unit No.	W_0	I_{D0}	g_{m0}	R_0	R_s	f (300°K)
	Volts	*ma*	*ma/v*	Ω	Ω	*mc/s*
20	35	6.5	0.7	340	300	5
23	55	40	1.0	310	200	—
24	70	35	1.0	360	140	20
27	50	120	5.0	—	—	—
29	40	20	1.0	300	500	17
30	37	20	1.3	500	160	18
32	40	40	1.6	200	152	50
35	28	15	1.0	110	36	31
36	20	12	1.0	—	—	—

response up to these frequencies would be a long and difficult experiment and was not attempted. However, an estimate of the response can be obtained from the performance of a unit when operated as an oscillator. It is shown in Appendix 2 that the maximum frequency at which a closely-coupled feedback oscillator can be made to oscillate is approximately equal to the cut-off frequency f_1 determined by the theory. The circuit used in the experiment is shown in Fig. 12. Close coupling for the feedback was obtained by winding one coil inside the other. The presence of oscillation was detected by a loosely-coupled radio receiver. Coarse changes in frequency were made by changing coils, while fine control was obtained with the variable capacitance K. The experiment was started by making the circuit oscillate at a comparatively low frequency, say 100 kc/s. The frequency was then increased continuously until no oscillation could be excited. It was also found possible to modulate the oscillation in the circuit of Fig. 13.

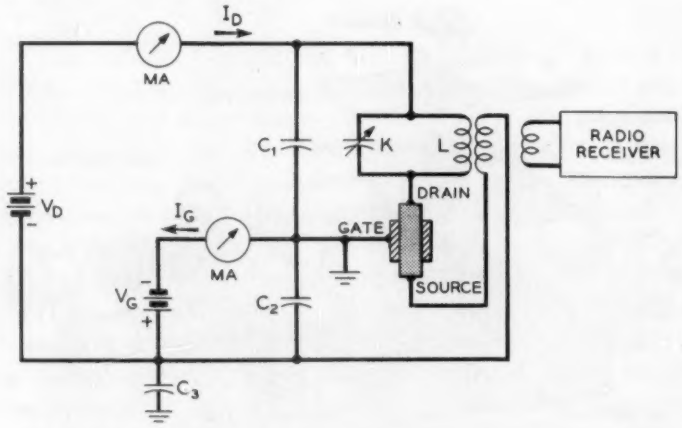


Fig. 12 — Test circuit used in estimating the frequency response.

3.0 *Experimental Results*

In all, test were made on nine specimens and the properties of these are shown in Table II. In these specimens the gate length, L, was about 7 mils, so that the pinch-off voltage at which the average field became equal to the critical field, E_c, was 20 volts. It is seen that most of the specimens had pinch-off voltages much in excess of 20 volts and would therefore not obey the constant-mobility theory. We shall describe one such unit, No. 32, in detail in Sections 3.3 and 3.4 below. However, two specimens, No. 35 and No. 36 had lower pinch-off voltages and they should approximately agree with the constant-mobility theory. Of these two, No. 35 was investigated more thoroughly and we will begin by discussing its behavior.

3.1 *Static Characteristics of No. 35*

The drain characteristics of unit No. 35 are shown in Fig. 14, and the gate characteristics in Fig. 15. It is seen that the unit had a pinch-off voltage of about 27.5 volts. With a gate length of 7 mils this gives an average field in the channel of about 1,500 volts/cm. This is 50 per cent higher than the critical field. However the square-root theory shows that no significant correction need be applied until the unit is operating much farther into the square-root range. In Fig. 16 the experimental points for $V_G = 0$ are replotted to the normalized scales of V_D/W_0 and I_D/I_{D0}. On the same figure are plotted the theoretical curves for the constant-mobility and square root cases. It is seen that the experimental curve is in close agreement with that for the constant-mobility theory.

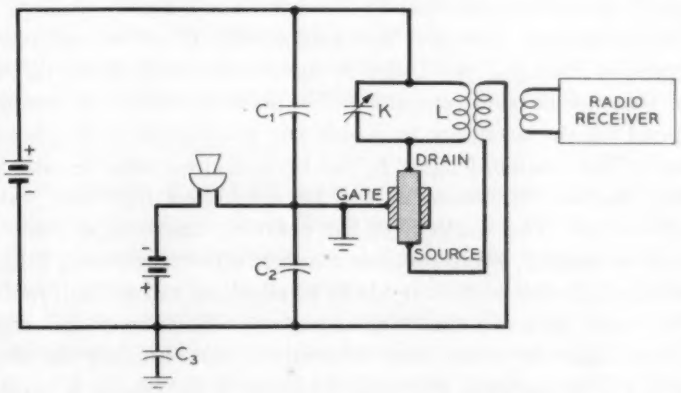


Fig. 13 — Modified circuit for demonstrating both amplitude and frequency modulation.

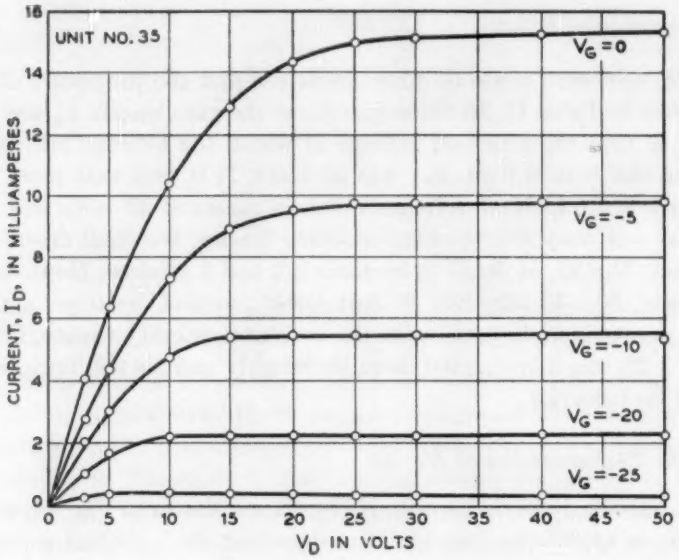


Fig. 14 — Experimental drain characteristics for unit No. 35.

In Fig. 17 a comparison has been made between the theoretical and the observed dependance of pinch-off current on the "effective gate bias," V_{Ge}, which differs from the applied bias by the voltage drop in the source resistor. The experimental points are shown as circles and the solid line is the constant-mobility curve adjusted to pass through the $V_G = 0$ point (i.e., $V_{Ge} = 0.55$). It is seen that there is good agreement.

Fig. 16 and 17 just described show that the current-voltage relationships are of the form predicted by the constant-mobility theory. It remains to determine whether the magnitudes of such characteristic parameters as I_{D0}, g_{m0}, etc., are in agreement with those calculated from the dimensions of the specimen. The accuracy of such a comparison will depend on the accuracy to which the dimensions of the specimen are known. The channel length, L, can be measured under a microscope while the channel thickness, a, can be calculated from the value of pinch-off voltage. The width Z of the channel, however, is much more difficult to determine. From sections made of other specimens, it appears that the channel cross-section tends to be elliptical rather than rectangular.* The exact influence that this departure will have on the effective width of the channel has not been determined, but certainly the effective value of Z will be less than the total thickness of the blank. A reasonable

* This effect can perhaps be minimized by the use of the 111 plane orientation of the original material during alloying.

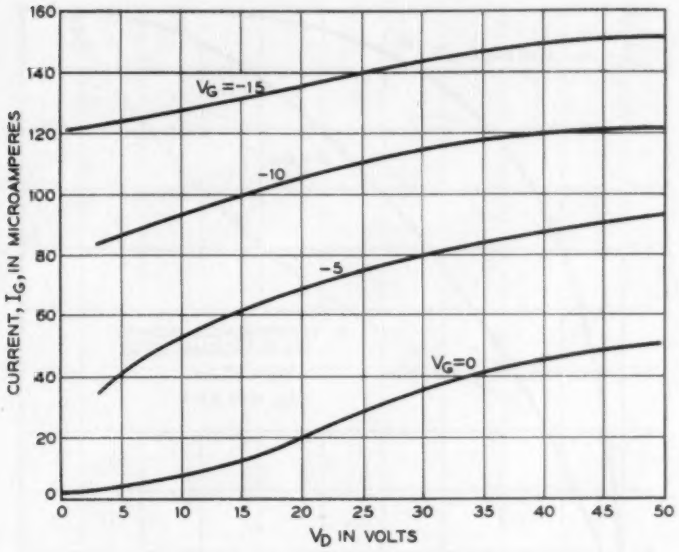


Fig. 15 — Gate characteristic for unit No. 35.

approach is to take Z as an unknown and on a basis of comparison of theory and experiment, determine its value and see if it is reasonable.

The measured transconductance at zero effective gate bias is 1.65 ma/v, whereas theory gives a value of 12 ma/v per cm in the Z direction. This comparison would suggest an effective channel width Z of 0.14 cm as compared to a blank width of 0.32 cm. Taking Z as 0.14 cm, the theory gives a value of pinched-off current at zero gate bias of 15.1 ma, and this is in good agreement with the experimental value of 15.8 ma. We conclude that the experimental values are in good agreement with the constant-mobility theory if Z is taken to be about one-half the blank thickness. This is deemed to be a reasonable supposition.

3.2 *High Frequency Performance No. 35*

The performance of the unit as an oscillator gave further confirmation of the theory. At comparatively low frequency, say 100 kc/s, the occurrence of oscillation was accompanied by a forward current at the gate. This is analogous to the grid current observed when a vacuum tube oscillates. When a negative bias was applied to the gate, the amplitude of oscillation (as observed on an oscilloscope) increased and, by adjusting the gate bias, a maximum peak to peak amplitude equal to twice the drain bias was obtained. This behavior shows that the transconductance was sufficiently high so that the amplitude of oscillation was not limited

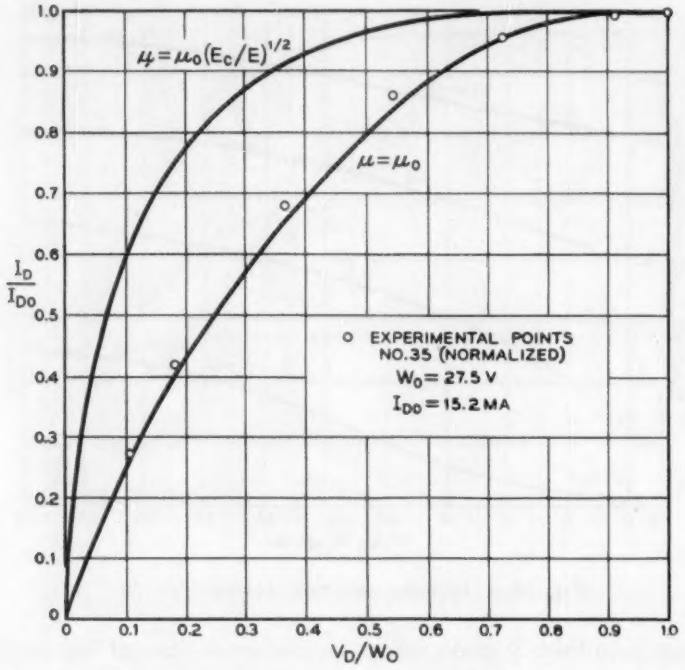


Fig. 16 — Comparison of the experimental drain characteristic for unit No. 35 with the two theoretical curves. It can be seen that the experiment points agree with the constant-mobility theory.

by it, but rather by the bias voltages. At higher frequencies the forward gate-current was not observed. Furthermore the amplitude of the oscillation was much less than twice the drain bias and decreased with frequency. This implies that at these higher frequencies the circuit had become transconductance limited. As the frequency increases so does the required transconductance and, since the maximum transconductance is obtained at very small bias, the gate bias and therefore signal amplitude must decrease.

The highest frequency at which continuous oscillation could be obtained was 31 mc/s. This figure agrees reasonably well with the value of 48 mc/s predicted from the theory. At slightly higher frequency, oscillation occurred for a few seconds after the bias voltages were applied, and then died out. This effect can be attributed to the temperature dependence of the transconductance. When the power was first turned on the transconductance was just big enough to support oscillation. After switch-on the temperature of the specimen increased slightly causing a decrease in conductivity of the germanium. Since the transconductance

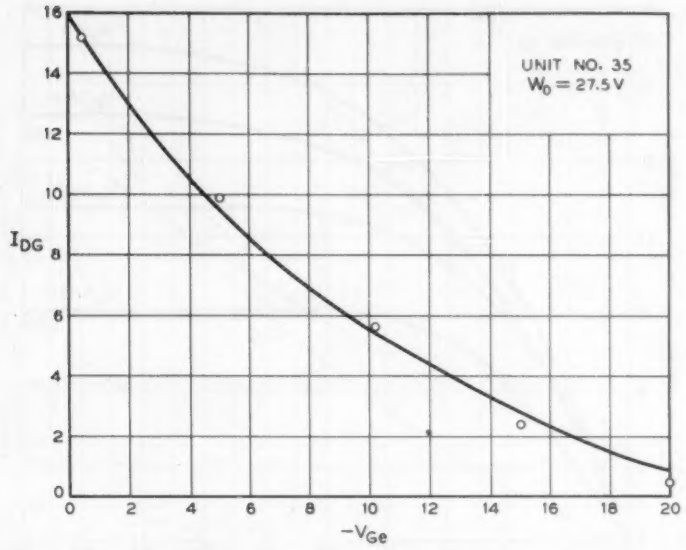


Fig. 17 — Dependence of saturated drain current on gate voltage. The solid line is derived from the constant-mobility theory while points are experimental values for unit No. 35.

is proportional to the conductivity, the transconductance fell slightly and oscillation ceased.

With this specimen it was found possible to amplitude and frequency modulate the oscillation. A telephone transmitter was used to modulate the gate bias as shown in Fig. 13. The mechanism for modulation is as follows. When the gate bias is varied both the transconductance of the device and its input capacity vary. The first effect causes amplitude modulation of the oscillations under conditions when operation is transconductance limited, that is, at the higher frequencies. Since the input impedance appears across the tank circuit, the change in input capacity causes frequency modulation under conditions when the input capacity and the tank-circuit capacity are of the same order of magnitude, that is, at the higher frequencies. It was found that both *AM* and *FM* could be obtained at frequencies close to 31 mc/s. At much lower frequencies only very slight *AM* was observed. It was not possible to look for *FM* at these lower frequencies, as the receiver was not equipped for *FM* detection in this range.

3.3 *Static Characteristics of Unit No. 32*

The static drain characteristics of unit No. 32 are shown in Fig. 18. It is seen that the apparent pinch-off voltage is about 40 volts. With a

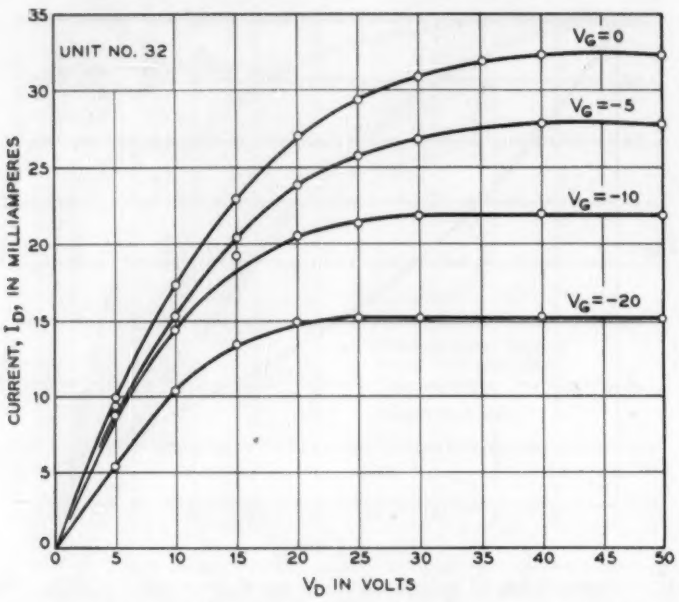


Fig. 18 — Experimental drain characteristics for unit No. 32.

gate length of 0.007″, an average field is obtained in the channel of 2300 volts/cm. This is well above the critical field and we should therefore expect departure from the constant-mobility theory.

The dependence of pinched-off drain current on effective gate voltage for unit No. 32 is shown in Fig. 19. The observed points are shown as circles. The two dashed curves are derived from the constant-mobility theory for two values of pinch-off voltage — $W_0 = 40$v and $W_0 = 55$v. It can be seen that the slopes of these curves are too steep to fit the observed points. (Compare their fit with that of Fig. 17). The *solid* line is derived from the square-root mobility theory for $W_0 = 55$v, and fits the observed points well. The value of 55 volts was chosen to give the best fit. Since the square-root mobility theory predicts a very gradual pinch-off it would be difficult to determine W_0 accurately from the $I_D - V_D$ characteristic, but referring to Fig. 18 one may see that a value of 55 volts is not unreasonable. Thus it seems that unit No. 32 follows the general form of the square-root mobility theory.

We shall now compare the *magnitudes* of I_{D0}, g_{m0}, etc., as determined by the two theories and the dimensions of the specimen, with the observed values. As before we assume that the effective width Z is such as to give agreement with I_{D0} and is reasonable. In Table III the observed values and values calculated from the two theories are compared. We compare four cases; A, B, C and D defined as follows:

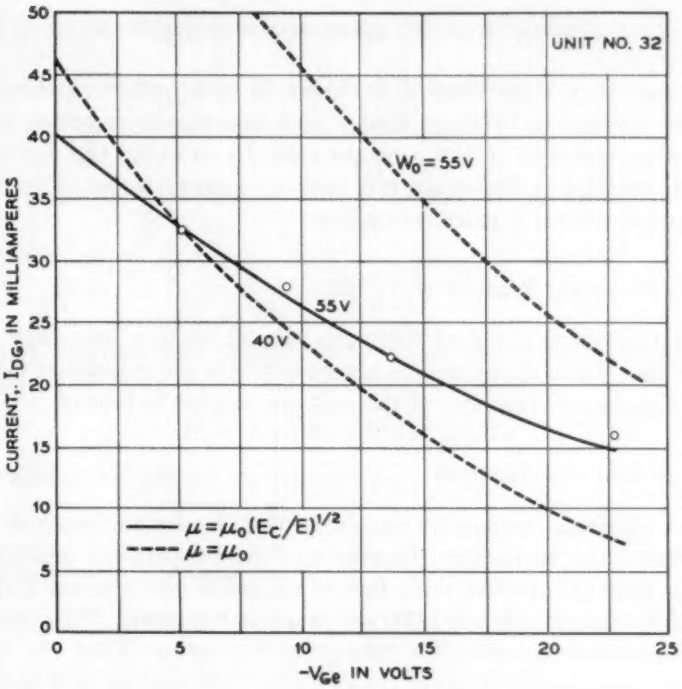


Fig. 19 — Dependence of saturated drain current on gate voltage for unit No. 32

TABLE III

	Constant μ theory		Square-root μ theory		
	A	*B*	*C*	*D*	Units
	(*obs*)	(*calc*)	(*obs*)	(*calc*)	
W_0 (assumed)	—	40	—	55	volts
W_0 (observed)	40	—	55?	—	volts
Z (assumed)	—	0.113	—	0.196	cm.
Z (observed)	0.32	—	0.32	—	cm.
g_{m0}	1.8	1.8	1.9	1.9	ma/v
I_{D0}	46	24	40	35	ma
f	50	70	50	46	mc/s

A. Observed values extrapolated to $V_{Ge} = 0$ (from $V_G = 0$ data) by using $W_0 = 40$v and the dependence of I_D and g_m on V_G given by the constant-mobility theory.

B. Values calculated from the $\mu = \mu_0$ theory using $W_0 = 40$v.

C. Observed values extrapolated to $V_{Ge} = 0$ (from $V_G = 0$ data) by using $W_0 = 55$v and the dependence of I_D and g_m on V_G given by the square-root mobility theory.

D. Values calculated from the square-root mobility theory using $W_0 = 55$v.

It is seen that, even when Z is chosen to give perfect agreement in g_{m0}, the agreement between theory and experiment is better in the square-root case than in the constant case. In addition, the value of Z that was assumed in the square-root case is a larger fraction of the blank width and therefore is more reasonable.

3.4 *High Frequency Behavior of No. 32*

Oscillations were obtained with unit No. 32 up to a frequency of 50 mc/s. This value is in reasonable agreement with the calculated value of 46 mc. The general behavior of the unit was similar to that of No. 35.

SUMMARY AND CONCLUSIONS

In the foregoing sections we have extended the design theory of field-effect transistors to include the effects of field-dependent mobility. It has been shown that when these factors are taken into account it should be possible to approach the 1,000 mc range at low power. Other possible advantages exist, however, for field-effect transistors. They can be designed for high power at lower frequencies. They can be used as direct replacements for pentode tubes without power supply alterations. It should also be possible, especially in silicon designs, to obtain very high input impedance and low noise, making the structure attractive for electrometer applications. On the debit side, the variation of g_m with V_G may lead to circuit complications in some applications. In the last analysis, of course, the eventual acceptance of the field-effect transistor hinges, as in so many cases, upon economic factors.

ACKNOWLEDGMENTS

The authors wish to acknowledge the valuable assistance of P. W. Foy and W. Wiegmann who fabricated the units described herein.

APPENDIX 1

It is of interest to calculate the transit time across the channel and the channel shape. In the constant-mobility case Shockley has shown[1] that the channel shape is

$$x = -(aI_0/I)\,[u^2/2 - u^3/3] = -6L[u^2/2 - u^3/3] \qquad \text{(A.1)}$$

where $u = b/a$. The field at any point $u(x)$ is given by

$$E(u) = I/g(u) = Ia/g_0 b = IaE_0/I_c b = W_0/3Lu \tag{A.2}$$

so that the transit time τ is given by

$$\tau = \int_0^L dx/\mu_0 E = 18L^2 \int_0^1 [u(u - u^2)/\mu_0 W_0]\, du \tag{A.3}$$

Upon integration

$$\tau = 3L^2/2\mu_0 W_0 \tag{A.4}$$

For the non-constant mobility case we can readily calculate the channel shape as follows. From (5.3) and (5.5) we write

$$I^2 = 4\sigma_0^2 b^2 E_c (dW/db)\,(db/dx) \tag{A.5}$$

The value of dW/db can be obtained from (5.1) and is

$$dW/db = 2W_0(a - b)/a^2 \tag{A.6}$$

When this value is substituted in (A.5) we obtain

$$I^2 dx = 8\sigma_0^2 E_c W_0 a^2 u^2 (1 - u) du \tag{A.7}$$

where again $u = b/a$. Upon integration we obtain

$$I^2 x = 8\sigma_0^2 E_c a^2 W_0 (u^3/3 - u^4/4) \tag{A.8}$$

Thus if we let $I = I_c$ when $x = L$, $u = 1$ we obtain

$$I_c = \sigma_0 a (2E_c W_0/3L)^{1/2} \tag{A.9}$$

as before. Using (A.9) to eliminate I from (A.8) we finally obtain

$$x = 12L(u^3/3 - u^4/4) \tag{A.10}$$

In Fig. 20 the channel shape for the constant and square-root mobility cases as obtained from (A.1) and (A.10) are shown. It can be seen that the two channels are of similar general shape and that the non-constant mobility case would behave experimentally like a constant-mobility channel with smaller a.

The transit time τ_c for the non-linear case is readily obtained from the integral

$$\begin{aligned} \tau_c = \int_0^L dx/\mu_0 (E_c E)^{1/2} &= (12L/\mu_0)(3L/E_c W_0)^{1/2} \int_0^1 (u^3 - u^4)\, du \\ &= (3L/5\mu_0)(6L/E_c W_0)^{1/2} = 1.47L/(\mu_0^2 E_c W_0/L)^{1/2} \end{aligned} \tag{A.11}$$

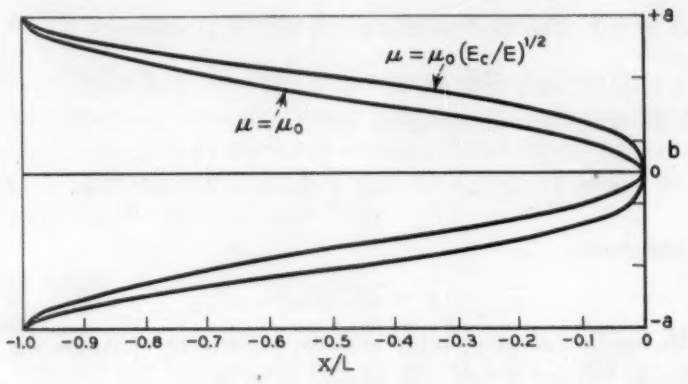


Fig. 20 — Comparison of theoretical channel-shapes for the constant and non-constant mobility cases.

The expression $(\mu_0^2 E_0/L)^{1/2}$ is the velocity corresponding to square root mobility in the average field $E_a = W_0/L$.

APPENDIX 2

Proof that in a unity coupled feedback oscillator circuit, the maximum oscillation frequency of a field-effect transistor is approximately equal to the theoretical value f_1.

We assume that the only reactive element in the transistor is the capacitance across the space-charge layer. This capacitance has to be charged through the channel resistance. Now if we represent the capacitance and resistance by lumped circuit elements, in accordance with Shockley's theory, the transistor may be considered to be as shown in Fig. 21(a). When an incremental voltage ΔV_1 is applied between gate and source, an incremental voltage ΔV_2 will appear across the capacitance C causing a change, ΔI, in drain current proportional to ΔV_2. That is

$$\Delta I \propto AV_2$$

Now at d-c, $\Delta V_2 = \Delta V_1$ and $\Delta I = g_{m0}\Delta V_1$, where g_{m0} is the maximum zero-bias transconductance. Therefore the constant of proportionality is g_{m0} and we may write

$$\Delta I = g_{m0}\Delta V_2$$

At any frequency $f = \omega/2\pi$

$$\Delta V_2 = \Delta V_1/(1 + j\omega CR)$$

$$\therefore \Delta I = g_{m0}\Delta V_1/(1 + j\omega CR)$$

and we may define an A.C. transconductance, g_ω, as

$$g_\omega = \Delta I/\Delta V_1 = g_{m0}/(1 + j\omega CR) \qquad \text{(A2.12)}$$

The input impedance at the gate is simply that of C and R in series Using this fact and the result of equation (A2.12), the transistor may be represented by the equivalent circuit of Fig. 21(b).

Now, consider the circuit of Fig. 22(a) which represents a feedback oscillator using a field-effect transistor. If we assume that the feedback coils are unity coupled, then oscillation will be just possible when V_1

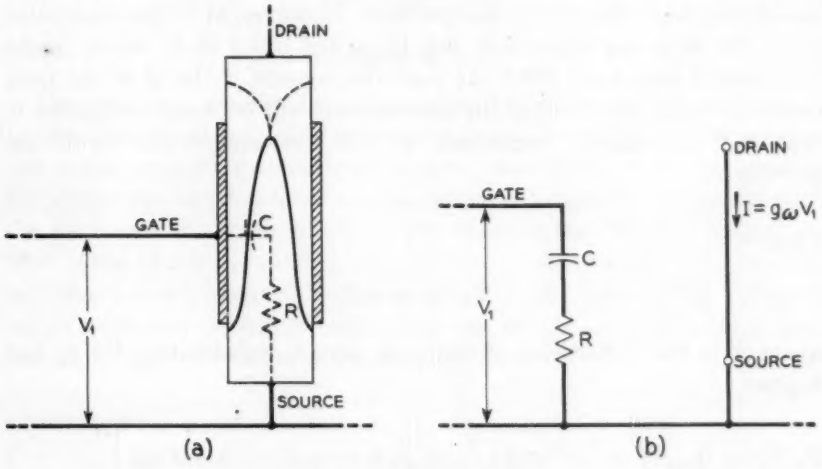


Fig. 21 — (a) Schematic diagram of equivalent circuit. (b) equivalent circuit.

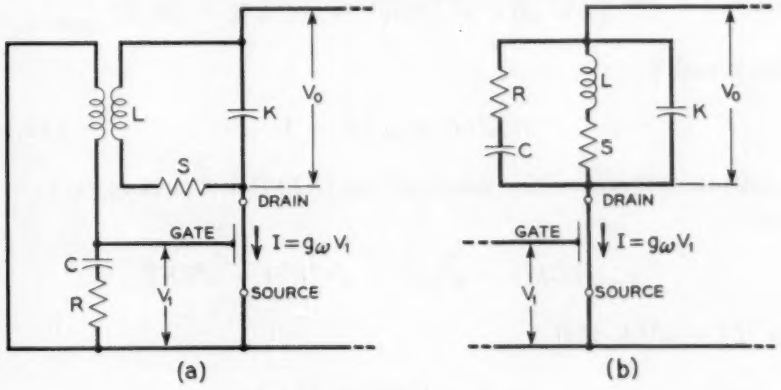


Fig. 22 — (a) oscillator equivalent circuit. (b) simplification of (a) assuming unity coupling.

and V_0 are equal in magnitude and phase. Furthermore, the input impedance of the gate will effectively appear across the tank circuit as shown in Fig. 22(b). Now the theoretical cut-off frequency of the transistor is f_1 where

$$f_1 = \frac{1}{2\pi CR} = \omega_1/2\pi \tag{A2.13}$$

At frequencies far below f_1, the shunting impedance of C and R will be high and can be neglected. If, then, the Q of the tank circuit (LKS) is sufficiently high, the circuit will oscillate. However, at frequencies close to f_1, the shunting impedance will be of the order of R, which in the units tested was about 100 Ω. At such frequencies, if the Q of the tank circuit is high, the effect of the resistance S will be small compared to that of R and may be neglected. We will now analyze the circuit neglecting S.

Then

$$V_0/V_1 = g_\omega Z$$

where Z is the impedance of the tank circuit. Substituting for g_ω and Z gives

$$\begin{aligned} V_0/V_1 &= [g_{m0}/(1 + j\omega CR)]\left[\frac{1}{j\omega L} + j\omega K + j\omega C/(1 + j\omega CR)\right]^{-1} \\ &= g_{m0}\frac{j\omega L[1 - \omega^2 LC - \omega^2 LK - j\omega CR(1 - \omega^2 LK)]}{(1 - \omega^2 LC - \omega^2 LK)^2 + \omega^2 C^2 R^2[1 - \omega^2 LK]^2} \end{aligned} \tag{A2.14}$$

This is real if

$$\omega^2 LC + \omega^2 LK = 1 \tag{A2.15}$$

Substituting this condition into equation (A2.14) and putting $V_0/V_1 = 1$ gives

$$g_{m0}\omega^2 LCR[1 - \omega^2 LK] = \omega^2 C^2 R^2[1 - \omega^2 LK]^2$$

and if $1 - \omega^2 LK \neq 0$

$$g_{m0}L = CR(1 - \omega^2 LK)$$

Substituting for $(1 - \omega^2 LK)$ from equation (A2.15) gives

$$g_{m0} = \omega^2 C^2 R^2$$

$$= \frac{\omega^2}{\omega_1^2} \cdot \frac{1}{R}$$

$$\therefore \omega^2 = \omega_1^2 g_{m0} R.$$

But

$$g_{m0} = \frac{1}{R_0} \div \frac{2}{R}. \tag{A2.16}$$

$$\therefore \omega = \sqrt{2}\omega_1 .$$

The maximum possible frequency of oscillation is approximately equal to the theoretical cut-off frequency as shown in (A2.16). This analysis was made neglecting the effects of stray capacities. However, each of the strays can be considered as equivalent to a capacity in parallel with the tank circuit and therefore simply changes the effective magnitude of K. Thus (A2.16) is valid in the presence of stray capacity. Therefore by using a closely-coupled feedback oscillator and determining the maximum frequency at which oscillations can be obtained the theoretical cut-off frequency can be approximately determined.

REFERENCES

1. W. Shockley, Proc. I.R.E., **40**, p. 1374, Nov., 1952.
2. G. C. Dacey and I. M. Ross, Proc. I.R.E., **41**, Aug., 1953.
3. E. J. Ryder and W. Shockley, Phys. Rev., **81**, p. 139, 1951; and W. Shockley, B. S. T. J. **30**, p. 990, 1951.
4. S. L. Miller, Avalanche Breakdown in Germanium. Phys. Rev. (To be published).

The Measurement of the Transient Power and Energy Dissipated in Closing Switch Contacts

By W. B. ELLWOOD

(Manuscript received July 29, 1955)

A new technique is described for the measurement of the power and energy dissipated in the contact gap of a glass-sealed reed relay (switch) on the closure of a special coaxial circuit of which the switch is a part. The method uses two cathode ray oscilloscopes with provisions made to study time intervals of from one microsecond to less than a millimicrosecond. Also, a brief resume is given of some experimental results on a few switches.*

INTRODUCTION

The loss or transfer of metal due to the electrical erosion of contacts used in the telephone system results in reduced life, increased maintenance, or the use of more costly contact metals. All this presents an important economic problem to the Bell System, promoting extensive study of contact phenomena.

In the ordinary uses of electrical contacts, the making and breaking of electrical circuits lead to discharges across the contacts, which may last up to the order of 500 microseconds. The behavior of some of these forms of discharge is fairly well understood† because their long duration and approximately constant voltage permit ready measurement of the effects and an analysis of causes and remedies for the behavior. As a result, these gross effects are under the broad control of the designer, who can allow them when the circuit application permits, but who can reduce them by "protection circuits" at increased cost. The use of sealed

* Recent Developments in Relays — Glass-Enclosed Reed Relay, W. B. Ellwood, Elect. Eng., **66**, pp. 1104–1106, Nov., 1947. Development of Reed Switches and Relays —O. M. Hovgaard and G. E. Perreault, B.S.T.J., **34**, pp. 309–333, March, 1955.

† Numerous papers by L. H. Germer, M. M. Atalla, and associates in Bell Telephone Laboratories.

reed switches is an alternative potential means of erosion control with other attractive features such as speed and reliability. As the use of these switches is extended to ever greater numbers of operations, the need increases for greater understanding and control of the various forms of discharge taking place therein.

The sealed reed switch has various aspects with more ramifications than switches operating in air. Once the contacts are enclosed in their own private atmosphere, the precious metals may be of less importance and the determination of the best combination of contact metals and gas atmosphere inside the tube becomes of considerable practical and economic interest. There is thus a need to understand the fundamentals of the process taking place in the switch both during the discharge and as a result of it. Since discharges with good switches involve times of a few microseconds only, significant measurements become more and more difficult as interest is extended into the transition and formative regions of the discharge.

Much of the experimental difficulty is due to the inseparability of switch and circuit. The life and behavior of the switch are dependent on the circuit which may have properties which are not well understood. Because of the short time intervals involved it is necessary to use circuits with known properties at high frequencies. In order to test the switch a standard circuit is needed. To study the phenomena in the switch the rest of the circuit should be a pure resistance, uncomplicated by energy storage elements. This requirement has not been fulfilled in previous contact studies but is solved by the coaxial cable circuit described here. This new circuit not only provides a practically pure resistance for the reed switch but also provides means of calibration, compensation, and cross-checking features, and can detect differences in time between events to ½ millimicrosecond. The uncompensated energy storage amounts to less than one per cent of the energy dissipated in a switch. Other circuits may give different effects with the contact but this circuit is a reference standard.

I. STATEMENT OF PROBLEM

A. The Circuit Aspects of the Switch Problem

A switch is always used as a part of a circuit, and the behavior of the switch is affected by this circuit as well as the current and voltages which it controls. The interchange of stored energy associated with the other circuit elements is modulated by the switch during the transition period in an at present unknown manner. Thus, the dissipation of energy in the

switch itself is determined by (1) the rate at which energy can be supplied to the switch from these storage elements and (2) the reaction of the various processes going on in the switch to the instantaneous current through it. The power dissipated in the contact region should provide important clues to the nature of the reactions taking place during the transition. The wear of the switch contacts depends on the electrical energy dissipated in the contact region and is known to be much greater than the wear of mechanical origin.

In order to obtain a true picture of the switch phenomena with minimum modification by circuit phenomena, it is necessary to study the switch in a circuit free from frequency sensitive and also other non-linear elements. This means the switch itself must be specially made to match the test circuit. The circuit used here influences the behavior of the contact in a simple manner instead of the usual complicated one. All data on switches included here is under one circuit condition, 150 ohms series resistance, which roughly approximates practical operating conditions.

B. The Switching Aspects of the Circuit Problem

In the usual texts discussing linear circuits including switches, the switching transition interval state is ignored. That is, the usual assumptions made in the boundary conditions eliminate the switch as a variable element. However, a certain amount of total circuit energy is always dissipated in the switch which becomes of great importance when the switch itself is under study. Furthermore, different mechanisms postulated to explain contact behavior lead to different circuit boundary conditions and it is impossible to choose between them without experimental knowledge of the current through the contact and the voltage drop across it as functions of time during the transition period.

C. The Measurement Problem

The experimental problem is to operate a test switch in a pure resistance circuit with a known battery voltage and to observe with a cathode ray oscilloscope (CRO) the instantaneous values of current and the corresponding voltage drops across the switch during the transition interval. From these observations, the instantaneous values of power and the total energy expended between the switch terminals can be calculated. This measurement must take account of the following:

(1) The non-linear nature of the contact (during transition) as a circuit element makes practical circuit analysis extremely difficult,

since the usual impedance concepts do not apply and the voltage drop and current must be determined by *separate but simultaneous* measurements. By non-linear is meant a condition where the voltage drop is not simply proportional to the current, its integral or its derivatives. In theory, the current and voltage can be calculated for any specific circuit by means of non-linear differential equations expressing the instantaneous reactions of the different circuit elements as functions of the current and the time. The resulting equations are often impossible of solution in a practically useful form and each one is a special case. In the case of the contact it is necessary to find an empirical expression for the voltage reaction as a function of both current and time before an analytical solution of the transition circuit can be attempted.

In the case of the coaxial cable, a measurement of voltage alone suffices to determine both voltage and current in the cable as the surge impedance of the cable is known and substantially independent of frequency. In the circuit described here the cable, the contact, the CRO and the battery constitute a form of ohmmeter in which the instantaneous voltage drop across the contact is compared with the corresponding drop across the cable with the same instantaneous value of current. The circuit aspects of the combination thus yield a number which is a ratio of resistances whatever the physical processes taking place in the contact.

(2) The shortness of the transition time (of the order of 10^{-7} to 10^{-10} seconds or possibly less) corresponds to extremely high frequencies so that the reactance of even one cm of wire is not negligible, neither are the admittances to ground.

(3) The non-repetitive behavior of practical contacts makes it necessary to observe single transients.

(4) Transmission phenomena ordinarily complicate both the phenomena to be observed and the interpretation of the observations. Some sort of parallel connections have to be made between the test circuit of which the switch under study is a part and the CRO which is used to measure the voltage and currents. The attached wiring constitutes a transmission line or lines, which provide delay times comparable to the time intervals under study and reflections and attenuation of the wave forms with possible mutual couplings.

These fundamental difficulties have long prevented an easy access to the basic information on the physical behavior of the contact under non-reactive circuit conditions. The circuitry of the deflection plate system, the transit times of the electron beam, and the photography of the screen are incidental minor problems.

D. Advantages of Present Scheme

The contribution of the technique of the present article is that advantage is taken of well known transmission phenomena in a coaxial line subjected to a voltage step to:

(1) provide an essentially resistive circuit* in which the contact can perform;

(2) reduce the number of possible interactions between the CRO and the contact including complications from parasitic circuits and mutual couplings between such circuits (done by shunting the reactive CRO circuit with the low resistance cable);

(3) take advantage of transmission phenomena to get electrically closer to the contact† than ever before; and to divide the pattern on the CRO screen into two areas: (a) the part relating to the contact behavior with negligible modification by the reflection, attenuation, and frequency dispersion effects taking place in the circuit; and (b) the other, showing same information *significantly* modified by the attenuation, reflection, and frequency dispersion effects of the circuit; and

(4) use of two CRO tubes at different points along the length of the coaxial line provides cross-checking features and a means of separation of contact phenomena from other phenomena in the circuit.

II. THE BASIC MEASUREMENT SCHEME

There are four basic elements involved in such measurements: (1) The development of a precision CRO beyond the capabilities of those

* The ratio of reactance to resistance of this circuit is calculated to have maximum values in the neighborhood of 200 and 600 megacycles and is less than 0.15. The reactive term is due to the reactance of the CRO in parallel with the coaxial line.

† The term "closer to the contact" needs interpretation. Any contact is usually connected to the circuit by leads which have some inductance, resistance, mutual capacitance, and capacitance to ground. All these provide local parasitic energy storage reservoirs which complicate the circuit analysis. In the case of the coaxial transmission line every unit length of center conductor has an inductance which is associated with an equivalent capacitance to the sheath so that an infinitely long line acts as a resistance, independent of frequency, and is dependent only on the geometry of the conductors and the specific capacitance of the dielectric between the sheath and center wire. In the coaxial switch structure used in this work the switch and its leads are surrounded with a conducting sheath constructed to have the same value of capacitance and inductance per unit length as the coaxial line into which it is incorporated. Thus the switch acts as it would if in a pure resistance circuit (equal to the surge impedance of the line) without effects due to lead inductance or capacity. The voltages observed on the line, a short distance from the contact are the same as those which would be observed at the contact with infinitely short leads (without inductance or capacity) but appear at the point of observation after a delay time of d/V_T seconds where V_T is the propagation velocity of the line and d is the distance.

available, with associated controls whereby the operation of the CRO is sufficiently synchronized with the contact under study, or vice versa, so that the phenomena in question appear on the CRO face at the proper instant and are photographically recorded; (2) The experimental development of the fundamental measurement circuit in which the contact operates and which has the properties outlined above; (3) The simple analysis (illustrated in Fig. 5) whereby the CRO record can be interpreted in terms of the basic parameters, voltage, current, power, and time without the need for complicated calculation; and (4) The experimental calibration of the CRO and proof of the record.

A. The Oscilloscope and Control Circuitry

Measurement of single transients involving time intervals of the order of nanoseconds* places special requirements on the CRO itself and the technique of operating it. The use of amplifiers at these frequencies introduces many circuit problems which are best avoided by using the CRO plates directly. The voltage sensitivity of the CRO tubes used permitted this study to be made in a useful range of voltage without need for amplification.

There are two CRO tubes used in these measurements — one (CRO-1) to record voltage† as a function of time, the other (CRO-2) to record voltage as a function of charge from which the energy dissipated can be calculated directly. There are also incidentally two cameras and counters associated with them so that corresponding photographs will be identified by the same number. The action of the oscilloscope control circuits will be described in Appendix A. A block diagram of both the oscilloscope control circuit and the coaxial line contact test circuit is given in Fig. 1. A schematic diagram of the arrangement of camera, counter and CRO tube face is shown in Fig. 2.

B. The Coaxial Cable Switch Test Circuits (Figs. 1, 3 and 4)

(1) *The Test Switch*

The glass sealed reed switch comprises a pair of contact reed springs of magnetic material sealed in opposite ends of a glass tube filled with an appropriate atmosphere. The normal gap between the contact ends of

* Unit of time equal to a millimicrosecond (American Standard Definition of Electrical Terms, Revised 1955). The abbreviation *ns* will be used hereafter for millimicroseconds. Also, from the same reference the term picofarad (*pf*) will be used as equal to micromicrofarad.

† The CRO tube is a voltage operated device. In the coaxial line both voltage and current waves coexist and are related to the characteristic impedance. One CRO reading defines both.

the springs inside the glass may be closed to make contact by application of a magnetic field along their axes. The distance between the contact spring end and where it is available for connection outside the glass is about three centimeters. This has an appreciable inductance at the frequencies which are involved in the closure of the contact. It was desired to test switches as manufactured without modifications such as auxiliary electrical connections through the glass. To do this the switch

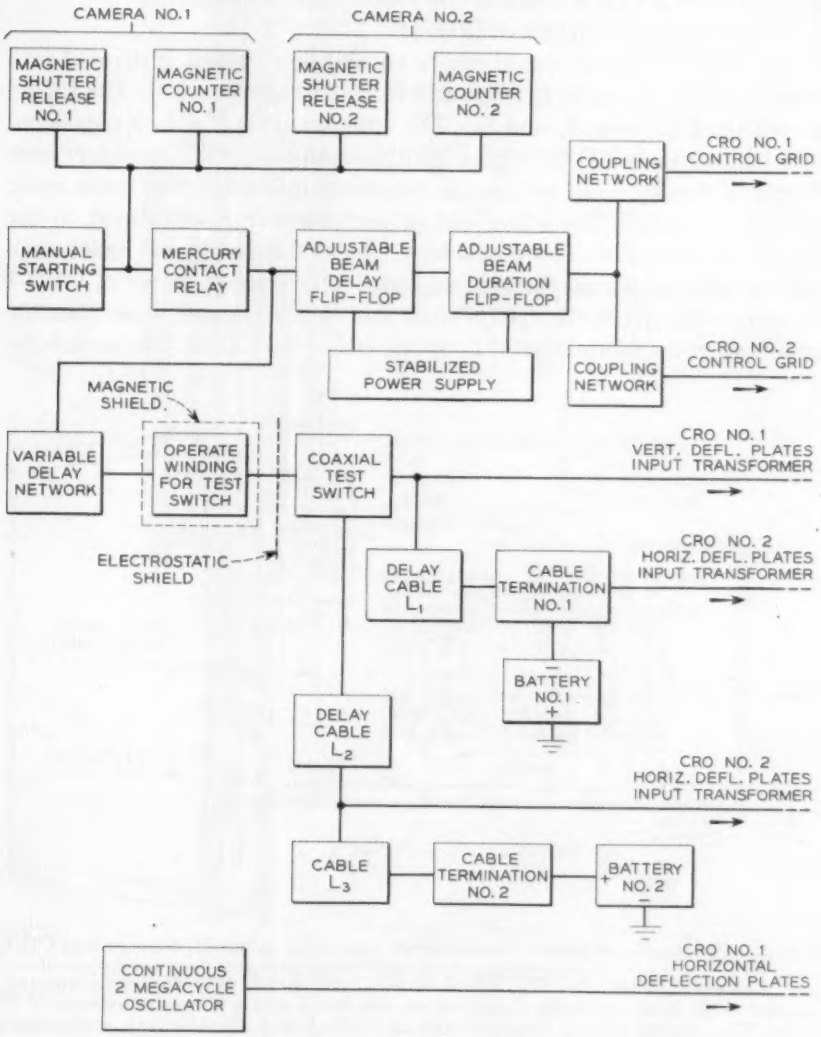


Fig. 1 — Block diagram of control and test circuits.

(Fig. 3) is supported inside a metal tube with an internal diameter so chosen that the switch elements and the tube combined form a length of coaxial conductor indistinguishable in surge impedance from the rest of the coaxial cable which is attached at each end. The operating winding of the switch is wound on the outside of the tube and outside the winding is a shell of permalloy to shield the nearby oscilloscope tube from the stray magnetic field of the switch and winding. This structure will be referred to as a COLS (coaxial line switch) for convenience.

(2) *The Contact Test Circuit (Fig. 4)*

The COLS is incorporated into a coaxial line* which is divided into three sections of length L_1, L_2, and L_3 feet. (Figs. 1 and 4). The COLS is connected between L_1 and L_2. The inner conductor of L_1 is connected to a battery (− pole) through a high resistance R_1 while the inner conductor of the $L_2 + L_3$ sections is connected in similar way to an *equal* battery (+ pole). The other end of each battery is connected to the sheath of the cable. The contacts of the COLS thus have twice each battery voltage across them when open. The vertical plates of CRO-1 (in series with R_3) are bridged with as short leads as possible between the inner and outer conductors of L_1 at one end of the COLS. The horizontal

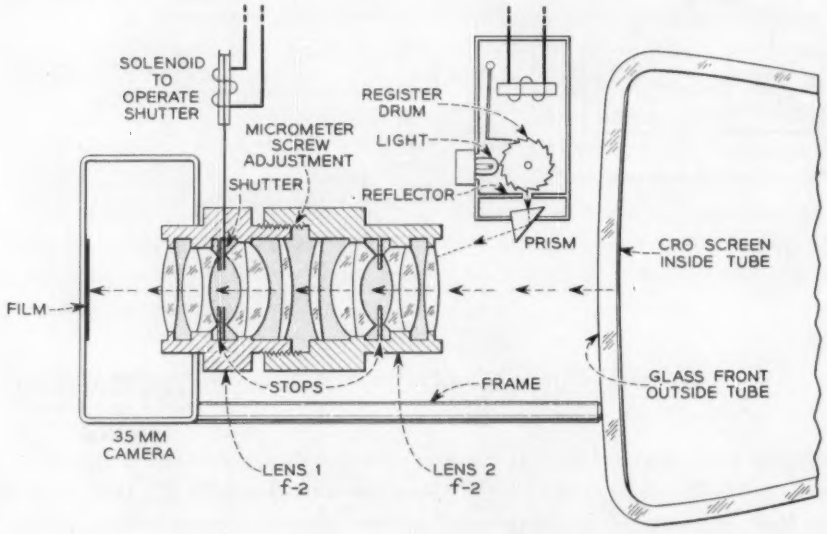


Fig. 2 — Schematic of camera lens system, operating solenoid, counter and CRO.

* Western Electric No. 724, which is a copper cored, polyethylene insulated, coaxial cable with a double sheath of copper braid and a surge impedance of 75 ohms. The double sheath is important as single braid sheaths leak appreciable energy to adjacent circuits.

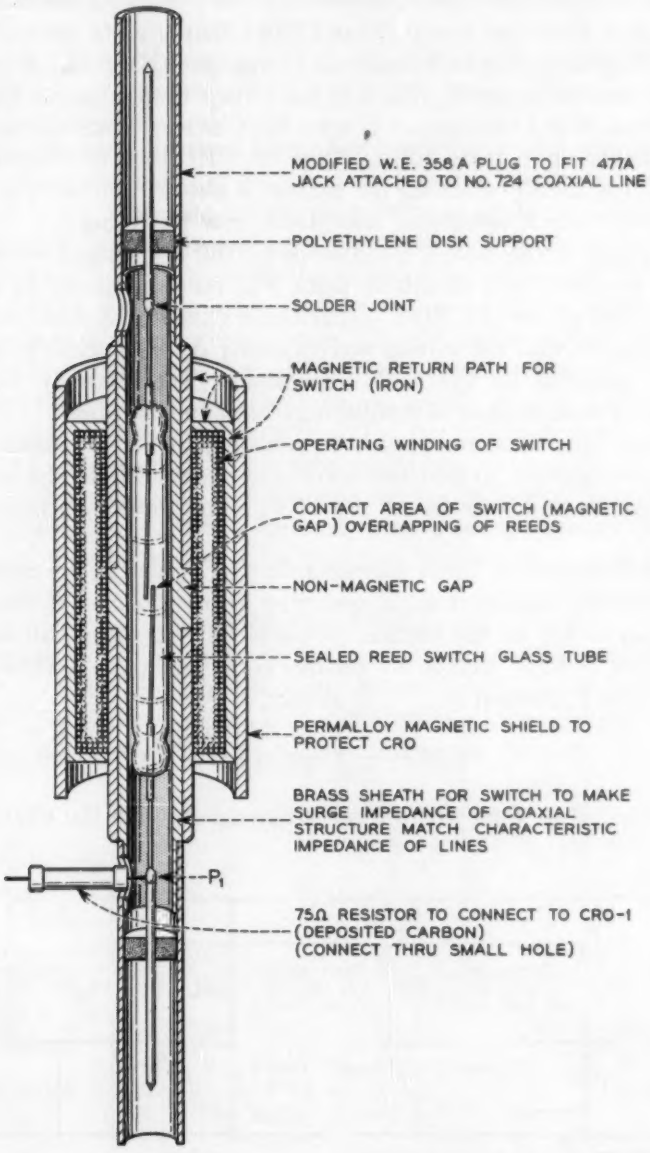


Fig. 3 — Schematic diagram of coaxial line switch.

plates $S - S'$ of CRO-1 are connected to a two megacycle oscillator which serves as a sinusoidal sweep. Thus CRO-1 simply plots the voltage between the center wire and sheath of L_1 near the COLS as a function of time on a sinusoidal scale. This is ½ the voltage drop (e) across the COLS as it closes. When the contact is open the CRO traverses a straight line at the voltage of battery No. 1. When the CRO indicates a steady zero voltage it is usually assumed the contact is closed. The latter is subject to further study if reflections take place prior to closing.

The plates of the second oscilloscope CRO-2 are bridged between the center conductor and sheath at point P_2, the junction of L_2 and L_3. The vertical plates of CRO-2 (capacitance C_0) see the same voltage as CRO-1 except that the voltage arrives nearly L_2/V_T seconds later and is slightly modified by the cable attenuation. The quantity V_T is the velocity of transmission of a pulse impressed on the cable.

Section L_1 is terminated with a resistance R_2 and a capacitance C_1 and is bent around so that the horizontal plates of CRO-2 connect directly to the terminals $(A - A')$ of C_1. The total capacitance $C = C_1 + C_0$.

If the transmission times corresponding to L_1 and L_2 are made *equal* then for every voltage drop, e, occurring in the COLS as it closes (and appearing as $e/2$ on the vertical plates of CRO-2) there will be a corresponding voltage change dv on the horizontal plates. Thus CRO-2 plots $e/2$ as a function of

$$v = v_0 - \int_0^t dv \quad \text{(Fig. 6)}$$

with $dq = idt = C\,dv$ where i is the current through the COLS and q

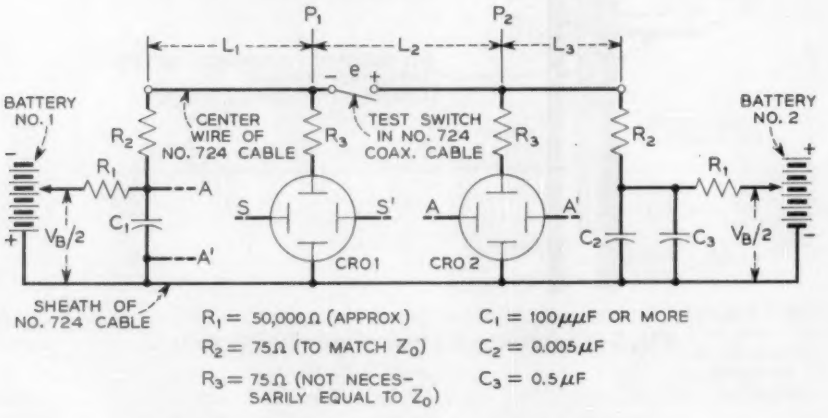


Fig. 4 — Coaxial line switch test circuit diagram.

the charge. Thus a typical vertical element of area dv wide and $e/2$ high is

$$\frac{e\,dv}{2} = \frac{e}{2}\frac{dq}{C} = \frac{ei\,dt}{2C} = \frac{P\,dt}{2C} = \frac{dw}{2C}$$

where P is the instantaneous power corresponding to the voltage e and w is the energy dissipated or stored. If the power is integrated* over the time from 0 to τ, (the time the switch requires to close) the total energy w dissipated on closing is obtained from CRO-2. Practically the upper limit of time is set up by the arrival of the reflected wave from the far end of L_1 which sets a limit on the time during which the coaxial cable may be regarded as a pure resistance. The COLS should then be closed ($e = 0$). However, in the case of some persistent discharges, this may not be true and the integration only applies to that portion of the closure process which occurs prior to the arrival of the reflected wave.

Section L_3 is long enough that reflections from that end do not return in time to interfere with observations on either scope. Nevertheless, such reflections should not be allowed to reverberate between the ends of the line. The far end of L_3 is therefore terminated in a dissipative network R_2 ($C_2 + C_3$).

(3) *The Oscilloscope Deflection Plate Circuits*

The resistance R_3 in series with the plates of the oscilloscopes plus the inductance of the lead wires both inside and outside the glass walls of the CRO tube and the capacitance of the scope plates constitute an r.f. coupling transformer,† (or impedance matching network) Fig. 7(a),

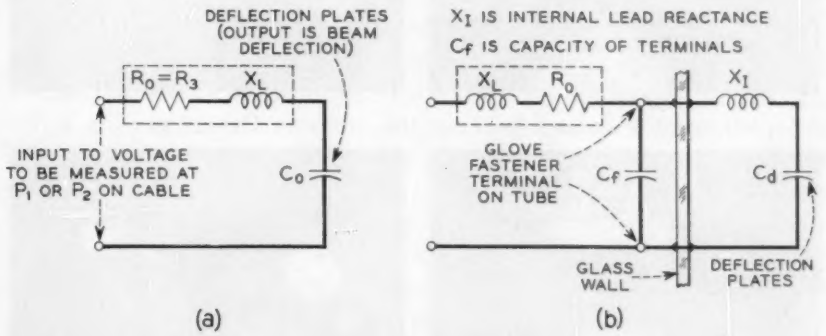
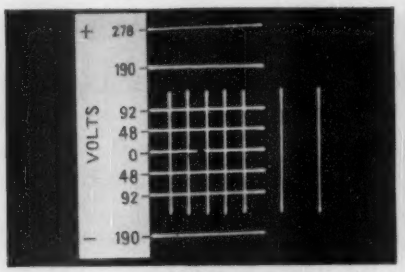


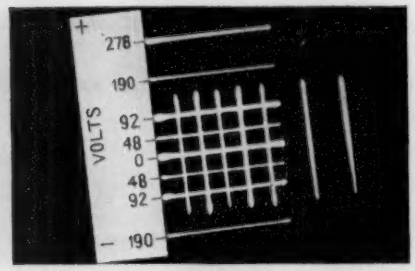
Fig. 7 — Vertical deflection plate circuits of CRO-1 and CRO-2. "Input transformer" or impedance matching network.

* A discussion of the errors of integration will be given later.

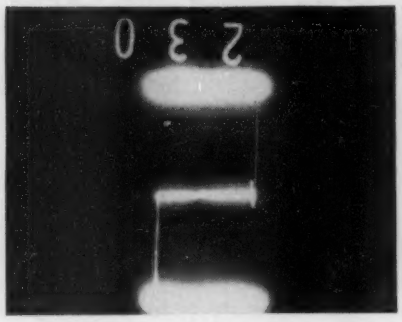
† Lee, R., Electronic Transformers and Circuits, 1st Ed. p. 123, 1947. Lewis, I. A. D., and Wells, F. H., Millimicrosecond Pulse Techniques, pp. 174–177, 1954.



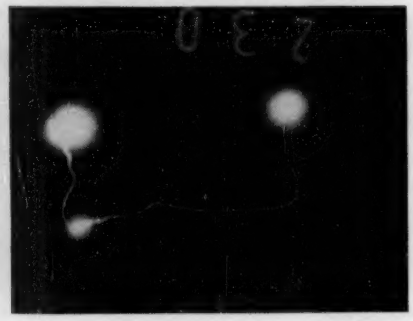
CRO-1



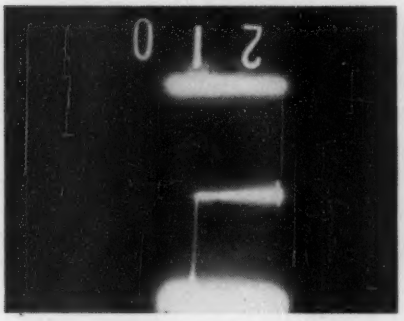
CRO-2



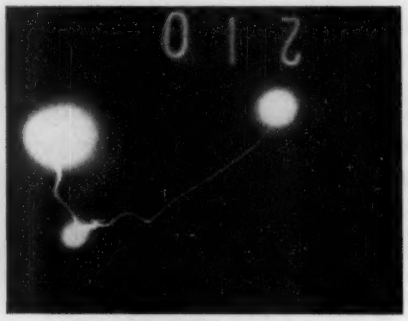
$L_1 - L_2 = 1$ ns



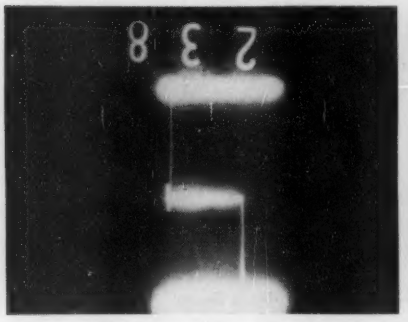
$C_0 = 4$ pf



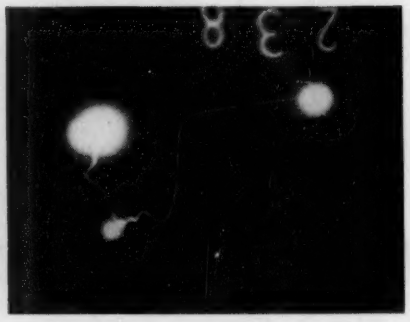
$L_1 = L_2$ ns



$C_0 = 4$ pf



$L_2 - L_1 = 1$ ns



$C_0 = 4$ pf

Plate A — Calibration and check data for CRO-1 and CRO-2.

which will be discussed (Appendix C) later. It is not to be confused with a simple RC circuit.

The CRO plates respond equally to all frequencies up to about 400 megacycles after which the response falls off rapidly. This means that an event taking place in a time of ¼ period or longer (0.6 ns) may be resolved on the screen and that time longer than this may be regarded as significant. Data presented in Plate A, Nos. 210, 230 and 238 (discussed later) may be taken as experimental confirmation of this time resolution.

III. THE INFORMATION DERIVABLE FROM THE OSCILLOSCOPE PATTERNS

A. Oscilloscope No. 1 (CRO-1) Voltage — Time Diagram, Fig. 5

This plots the cable voltage at point P_1 as a function of time on a sinusoidal scale. Point P_1 is as close to the contact as practical (within two inches of cable). A typical record is shown schematically in Fig. 5. The diagram has two important areas with special features.

(1) In between the *O* line (contact open) and the *C* line (contact closed) is a connecting line (*OVD*) with varying degrees of intensity depending on the velocity of the beam at the moment of recording. This line is the record of the voltage changes taking place both in the cable and in the contact as the current increases during the transition interval from open to closed. This is the *original* voltage change vs time data (*OVD*) — modified only by the attenuation of the two inches of coaxial line, the limitations of the coupling transformer and any electron transit time effects taking place between the oscilloscope plates.

(2) At the bottom of the CRO plate is a third horizontal line, the *R* line, caused by the reflection of the wave corresponding to the closed contact. In between the *C* line and the *R* line is a connecting line (*RVD*) which is the same data as above but modified by the effects of the cable attenuation and the energy absorption during reflection from the network at the end of L_1. This *RVD* line is considerably heavier than its *OVD* counterpart due to the slower CRO beam velocity resulting from the selective elimination of the higher frequency components in the cable attenuation and reflection processes. This *RVD* line provides additional useful data because (1) the effects of attenuation and reflection by various lengths of coaxial cable can be studied by variation of L_1 and L_2 and C_1 while observing the corresponding changes in the *RVD* line; (2) the behavior of the *RVD* line is often modified by reopening or reclosing of the contact (chatter) and can be used as an indicator thereof; and (3) with prolonged discharges the interaction of the reflected wave can be studied.

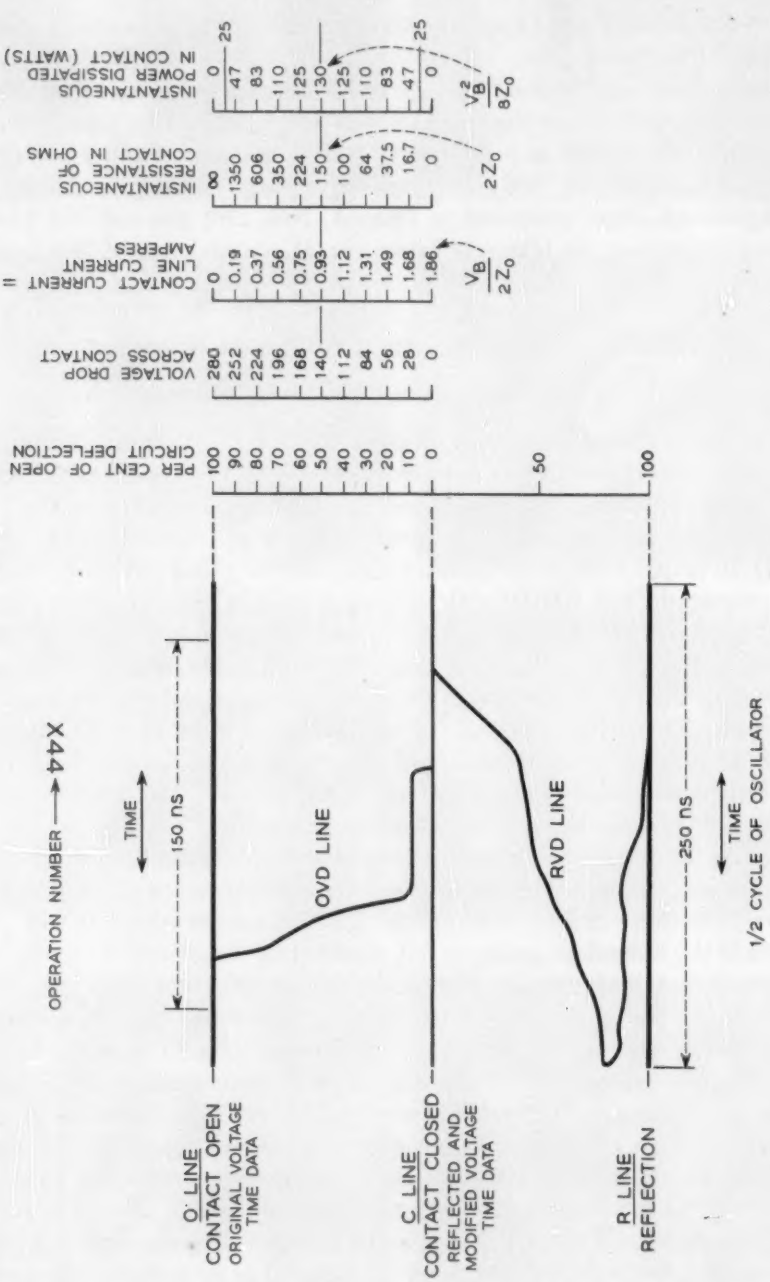


Fig. 5 — Data analysis of typical voltage time diagram for CRO-1 (voltage 280 R_0 = 150 ohms).

(3) Properties of the *OVD* line (Fig. 5*). The *OVD* line is terminated at the top by the *O* line and at the bottom by the *C* line. A scale of *voltage* reading from zero to the total battery voltage may be erected normal to the *C* line. A corresponding time scale may be provided along the *C* line. From these scales the voltage drop in the COLS at any time during the transition interval may be read directly. The voltage drop vs time is, however, not the only information obtainable from this *OVD* line. Since the cable is a pure resistance the corresponding concurrent values of current, power dissipated, and instantaneous resistance of the switch, as functions of time are all simply related to this voltage drop as shown in Fig. 5. The power scale is double valued with its maximum in the middle.

B. Oscilloscope No. 2 (CRO-2) Voltage Charge Diagram, Fig. 6

This provides additional and important information some of which is not available from the record of the first oscilloscope.

(1) It provides a greatly expanded non-linear time scale (about 15×) (Fig. 6) for the initial portion of the voltage drop record. This expanded scale is provided by the discharge of the condenser C_1 when the fall in voltage taking place in the contact reaches the far end of L_1. Means for (a) calibrating the non-linear time scale, and (b) checking the over-all transmission or recording characteristics of the oscilloscope and its coupling transformer, are provided by adjustment of the differential line lengths $(L_2 - L_1)$ in discrete steps corresponding to known time intervals.

(2) The pattern on CRO-2 (Fig. 6) is also divided into two areas corresponding to the *OVD* and the *RVD* lines of CRO-1 (Fig. 5). The demarkation is indicated by a sharp increase in voltage from the zero voltage base line and occurring at a time nearly $3T$ ns after the initiation of the voltage fall in the COLS contact: (T is the one-way transmission time of a coaxial segment $L_1 = L_2$) or $2T$ ns after the voltage starts to fall on CRO-2.

(3) The voltage corresponding to the voltage changes taking place in the COLS contact closure interval (vertical deflection) and recorded by CRO-2 may all be interpreted similarly to CRO-1 in terms of power, resistance and current. They are slightly modified by the attenuation of the line L_2 but this error can be determined by comparison of long and

* Fig. 5 shows a numerical example worked out for the highest voltage used, taken in round numbers as 280 volts. Values are worked out for voltage steps of 10 per cent. These quantities are all related through the surge impedance of the cable otherwise separate determinations of voltage and current as functions of time would be required.

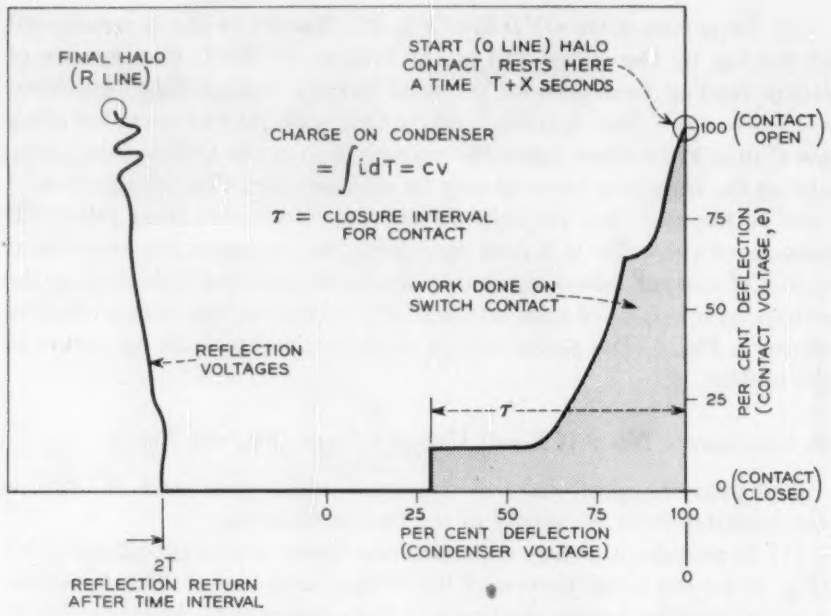


Fig. 6 — Data analysis of typical voltage charge diagram for CRO-2 (work diagram).

short lines. The steepness of the wave fronts shown in Plate A, Nos. 210, 230 and 238, show the error to be less than 10 per cent for the 100 ns coaxial lines and one ns time interval.

(4) The total area under the curve from zero to $2T$ represents the total energy dissipated from the beginning of the closure process up to $2T$ microseconds. For times greater than $2T$ the record has only qualitative significance but provides important clues as to the behavior of the contact at later times. The record may be complex because of the effect of reflections and thus require rather detailed study before conclusions can be drawn from it.

(5) A feature of CRO-2 is the ability to detect the flow of small currents at substantially constant voltage by using very small values of C_1. These currents may not be observable on CRO-1 as the voltage drop (iZo) across the cable may be too small. In the CRO-2 circuit the small currents produce a definite horizontal change $\Delta v = i\Delta t/C$.

C. Procedure for Equating Transmission Delay Times of L_1 and L_2

To do this one must consider the circuit with C_1 equal to the capacity (C_0) of the horizontal plates of CRO-2 alone (total $C = C_1 + C_0$).

Referring to Figs. 4, 6 and 8, the beam of CRO-2 remains at its initial position, corresponding to the voltage $V_B/2$ on the cable section L_2, until the COLS closes. This starting position is represented by the spot or halo at $-V_B/2$, $+V_B/2$ (shown reversed (as $+V_B/2$) in diagram to coincide with photograph). As soon as the COLS begins to close the voltage changes and a voltage wave propagates from the COLS in each direction. No voltage change with time takes place (that is the beam stands still producing the halo at the right) for either the horizontal or vertical plates of CRO-2 until this wave reaches them.

If L_2 is longer than L_1, Fig. 8(a), then the beam of CRO-2 remains at the battery voltage $V_B/2$ while it moves horizontally as the horizontal plates are discharged by the voltage pulse arriving at the end of L_1. If the difference between L_2 and L_1 is small, then the horizontal line will be terminated by a sharp drop in voltage corresponding to the arrival of the voltage wave at the end of L_2. Thereafter, the voltage on both the horizontal and vertical plates of CRO-2 will fall at a rate determined by their discharge time constants while the corresponding line drawn on the screen will start off at a steep angle toward the origin.

On the other hand, if L_1 is greater than L_2 then the vertical voltage

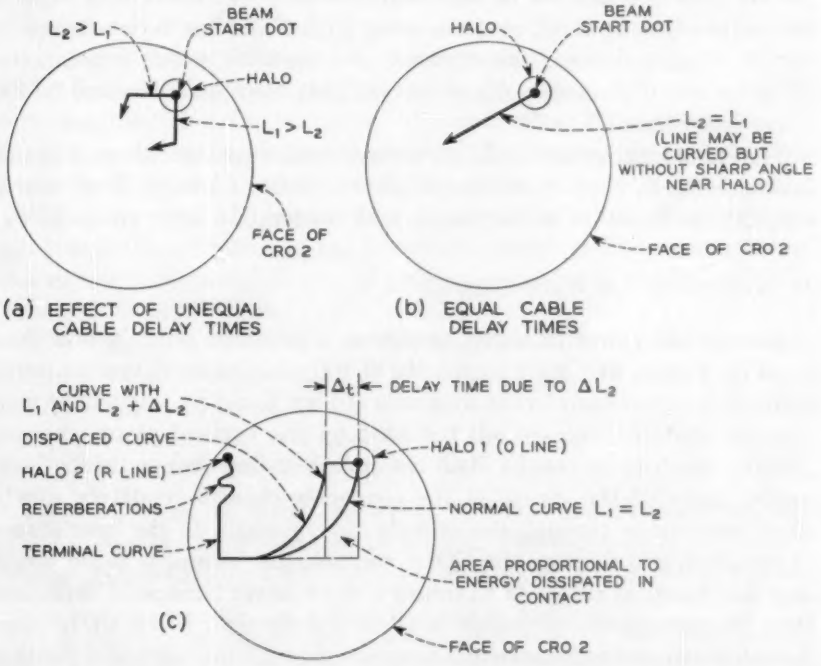


Fig. 8 — CRO patterns for equalizing cables and time calibration.

will fall before the horizontal voltage changes. Then the beam will move vertically down from its initial position until the L_1 wave has arrived at the end of L_1 when the combined effect is to move the beam at a roughly 45° angle.

If, however, L_1 and L_2 (Fig. 8(b)) are *exactly* equal the two waves arrive at the horizontal and vertical plates at the same time and the beam moves directly from the initial position toward the origin Plate A, No. 210. This condition will be referred to as the balanced condition. If the contact were linear or closed instantaneously and there were no parasitic oscillating circuits, the line traversed by the beam would be a straight line directed from the initial (open) position toward the origin. The end of this process may be indicated by a small spot corresponding to zero contact voltage from which there is a sudden rise in voltage corresponding to the arrival of the reflection of the closing contact condition from the end of L_1 ultimately ending in a second halo (at the left) corresponding to the R line of CRO-1.

Plate A, Nos. 230, 210 and 238, show the effects obtained for a one ns difference between L_1 and L_2 with $C_0 = 4$ *pf* only. It will be noted that for the balanced condition ($L_1 = L_2$), the theoretical straight line is departed from in a number of ways. One, initially the line is very slightly curved concave upward; two, the latter part of the line is terminated in certain wiggles showing the existence of a parasitic circuit which is excited by the impinging voltage waves. This time scale is roughly 400 times faster than CRO-1.

The cable lengths L_1 and L_2 are always made equal or balanced in this manner with C_1 equal to the scope plate capacity C_0 only. Now we will consider the behavior of the circuit with moderately large values of C_1.

D. Analysis of The Work Diagram

Assume the above balanced condition is provided initially and then C_1, Fig. 4 is set to a known capacity of 100 *pf* or more. A typical result is shown schematically in the diagrams of Figs. 6 and 8(c). It will be seen (Fig. 6) that the voltage fall recorded by the vertical plates changes initially much more rapidly than the corresponding change on the horizontal plates as the charge of the condenser changes relatively slowly since the current through the contact is very small. In the later stages of the discharge through the COLS contact, the current is much larger and the change in charge of C_1 causes a much larger horizontal deflection than the corresponding change in vertical deflection. If the COLS contact closes to zero resistance in a time less than $2T$ the vertical deflection

may drop to zero while the horizontal deflection of CRO-2 may still continue to increase. After a time interval $3T$ from the start of the discharge the reflected wave from the distant end of L_1 arrives at CRO-2. The sign of this wave is opposite to the sign of the wave causing the initial vertical deflection of CRO-2, causing the second vertical deflection of CRO-2 to increase from zero sharply and eventually to return to the initial voltage. This may be denoted by a second halo spot. The time difference between the starting dot on the CRO-2 and the upward break in voltage is thus $2T$ (in this case 200 ns) — the beam remaining at the starting point for a time T plus whatever time interval X occurred between the initiation of the beam and the start of the fall in contact voltage. During this time interval $2T$, the vertical deflection is proportional to the voltage drop across the COLS, and the horizontal deflection is proportional to $\int_0^t \frac{idt}{C}$.* The events taking place after the time $2T$ are not subject to this analysis for the cable L_1 can no longer be considered purely resistive for times greater than $2T$. However, contact events such as sudden arc extinctions or contact opens, taking place after this time interval may cause variations in the charge reaching C_1 and hence produce horizontal deviations in the slope of the line starting at $2T$. A horizontal change toward the starting dot indicates an opening of the contact, a change away from it, the termination of an arc or glow.

The *time intervals* corresponding to the horizontal deflections of the work diagram may be estimated by deliberate introduction into L_2 of additional lengths of line ΔL_2 with known times of traverse. This causes the beam to move horizontally a distance corresponding to the time of traverse of ΔL_2 before the vertical motion starts (Fig. 8(c)). In this manner a calibration of the horizontal charge scale as a function of time can be determined. A typical pattern with $\Delta L_2 = 10$ ns is shown in Plate B, No. 132. In Plate B, No. 594, $\Delta L_2 = 5$ ns.

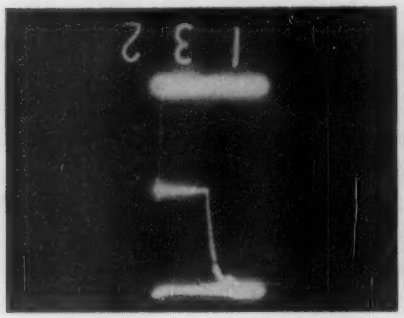
In the examples shown here all tests are made with the nominal length of L_1 or L_2 equal to a travel time of 100 ns and a ΔL_2 of 5 ns.† C_1 is 100 *pf* and $R_2 = Z_0 = 75$ ohms.

E. The Voltage Calibration Scales

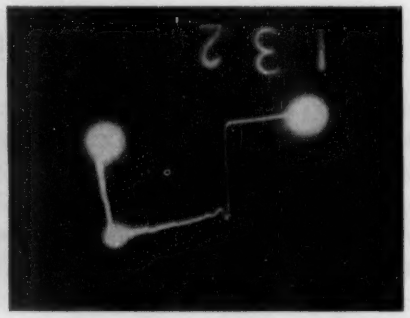
The conversion of CRO deflection to numerical values is an important part of the measurement process. For calibration purposes a voltage grid of the two CRO tubes was made and it is shown at the top

* The errors of integration will be discussed in Appendix C.

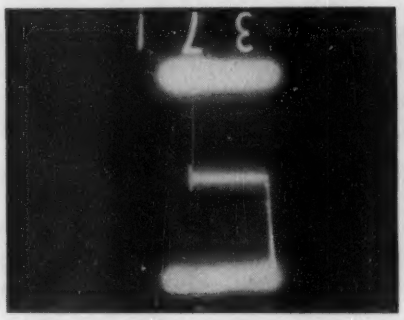
† For side-stepping the starting spot halo (Appendix C). This time interval may be insufficient to side step the halo for slow forming discharges like glows.



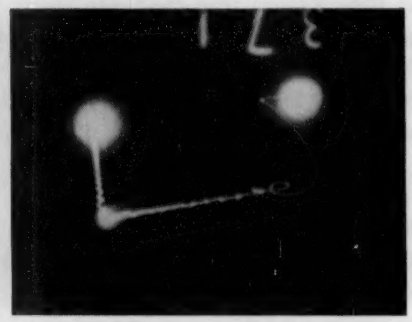
$L_2 - L_1 = 10$ ns



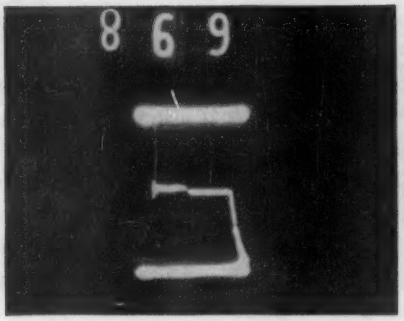
$C_0 + C_1 = 104$ pf

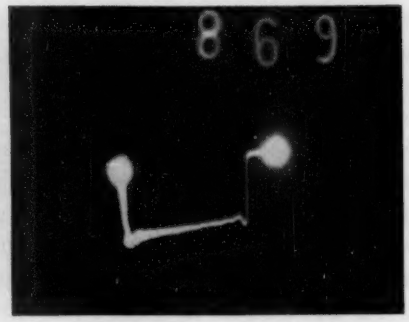


$L_2 - L_1 = 1$ ns

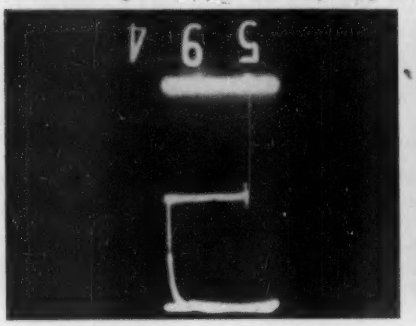


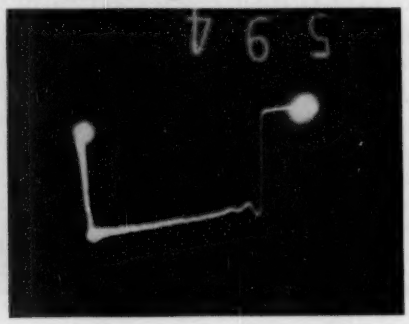
$C_0 + C_1 = 104$ pf





$V_B = 190$ volts — short arc 14 volts duration about 85 ns to metallic closure current 1.15 amps 16 watts 137 mmj. Typical initial arc voltage about 35 volts.





$V_B = 278$ volts. Drop to 14 volt arc. Initial arc voltage 63 volts. Drop in 2.5 ns at 200 ns arc voltage is still 24 volts, at 15 ns voltage is 32 volts. Very typical behavior every time. Switch A.

Plate B — Upper half, continuation of Plate A. Lower half, data on Switch A.

of Plate A. All voltages were measured by superimposing a transparent copy of this grid on the CRO pattern. The distances were measured to ± 0.01 inch on a three time enlargement which corresponds to ± 3 volts or 1 per cent of full scale. Since there is an error of positioning of about the same amount the combined error is about ± 5 volts. Variations less than this are of no significance. The six degree rotation of CRO-2 grid is a mechanical condition. For both CRO's the coordinate axes are perpendicular within experimental error. Also, the zero voltage line (horiz.) has a burned spot on the phosphor providing a fiducial point for zero contact voltage.

F. Effects After 2T Time

The two oscilloscopes at first see the *same* events displaced in time by the delay interval T. For times greater than $2T$, this simple condition is no longer true and CRO-1 and CRO-2 may give results that are seemingly quite contradictory. The behavior of the two scopes may be reconciled and understood, however, in terms of the reflections which take place both at the end of the line and at the contact gap. The CRO-2 originates no large reflections and gives a fairly true picture of the behavior of the switch from 0 to $2T$. The CRO-1 is located next to the contact but the voltage it measures after $2T$ ns depends on the state of the contact gap at the moment the reflected wave returns. If the gap is closed to zero resistance, the reflected wave is merely recorded and goes on past both CRO-1 and CRO-2 to be absorbed in the termination at the end of L_2. If the gap is open, or has a discharge (arc or glow) of some kind, the instantaneous resistance of the gap or discharge is added into the circuit and the contact gap becomes a discontinuity in the line causing the first reflected voltage to be partially transmitted through the switch and partially reflected. The amount and sign of the new voltages are determined by the ratio of gap resistance to the cable impedance. Thus from a detailed comparison of the reflected and transmitted voltage waves as measured by the two CRO tubes one can study the impedance of the arc and glow discharges still present at $2T$ time. This is a major contribution of this dual CRO and coaxial cable technique; however, application of it to specific cases is very complicated and is not discussed further.

IV. EXPERIMENTAL RESULTS ON SWITCHES

With the CRO circuit designs as shown in Figs. 7(b) and 9(b) (see Appendix C for discussion), the study of various switches was under-

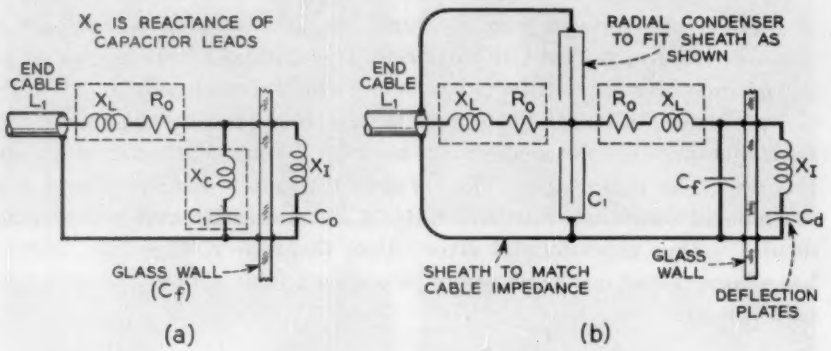


Fig. 9 — Horizontal deflection plate circuits for CRO-2.

taken. Some typical data on a few switches (listed in Table I) are presented to indicate the power of the measuring system to reveal the fine structure in typical discharges; and to illustrate a few typical phenomena associated with the closure transition. All the switches are of the glass enclosed reed type — differences being ascribed primarily to filling the tube with different gases. The surface of the contact area of the reeds is plated with 0.0001″ of gold which is diffused into the base metal (either No. 52 alloy* or 45 perminvar†) by heat treating in hydrogen for a con-

TABLE I

Switch	*A*	*C*	*K*	*E*
Metal	Perminvar	52 Alloy	Perminvar	Perminvar
Gas	97%N_2–3%H_2	Air	Helium	Hydrogen
MBV* Volts	240	327	210	292
Pressure ATM	1.7	1	1	1
Comment on behavior character of switch	Very fast, uniform	Slow, irregular	High voltage glows† frequent	High voltage glows, metallic closure at low voltage

* The minimum breakdown voltage (MBV). All values of MBV are from Meek and Crags Electrical Breakdown of Gases.

† The terms arc and glow are used to refer to discharges with voltage characteristics in the range usually encountered with discharges with similar names but taking place between fixed electrodes. The end of the arc can be identified by a distinct step in voltage.

---

* Driver Harris alloy of 50 to 51 per cent Ni balance Fe used to seal to Corning 0120 glass. Resistivity 43.2 microhm cm M.P. 1425°C. Thermal conductivity −183 watts/cm/°C; TS 70,000 to 15,000 p.s.i. sp.g. 8.247; coefficient of expansion at 20° to 100°C, 9.3 × 10^{-6}.

† Elmen, G. W., Perminvar — Alloy of 45% Nickel 25% Cobalt 30% Iron, Elec. Eng., **54**, pp. 1292–9, 1935.

trolled time and temperature. The reed assemblies are sealed in glass and the switches filled after only a short exposure to air.

The switches were tested under the four battery voltage conditions given in Table II with the corresponding values of current and peak power which are dissipated in the contact when it matches the combined cable impedances = 150 ohms.

Discussion of the Data

Table III shows some typical data obtained for the *four different switches* with a battery voltage of 278 volts, which condition brings out the greatest contrast in behavior. Table IV shows the behavior of *one switch at four different battery voltages.* In addition to the Tables, the following discussion is a brief resume of the effects observed for closing contacts.

(1) At low voltages (48 volts) the data shows extremes of behavior ranging from closure (to metallic contact) in one ns or less with no trace

TABLE II

Battery E.M.F., volts	Max. contact current, amperes	Max. power (P), dissipated watts
48	0.32	3.84
92	0.615	14.0
190	1.26	60.0
278	1.86	129.0

of an arc for the hydrogen switch, to the rather long drawn-out changes in voltage drop of the air filled switch taking almost ½ microsecond to go from 48 volts to zero.

(2) Other phenomena encountered were: (a) metallic closures with no arcing (Plate C, No. 851), (b) closures to the 14 volts "short arc" followed by metallic closure (Plate B, No. 698), and (c) metallic closure first (Plate C, No. 857), followed by arcing at 14 volts.

(3) In general as the battery voltage is increased the frequency of closures without arcing decreases and the number of arcs and their duration time increases. There are many rapid closures to the 13- to 14-volt arc condition which is substantially constant for its duration which may range from about 10 ns to over several hundred. It should be pointed out that the voltage across the contact is ultimately removed by the reflected wave. The initial drop to the 14 volts arc (Plate B, No. 698) may take place in a time less than the resolution of the scope which is about ½ ns.

(4) The initial rapid drop in voltage may not proceed directly to the 14-volt level at the higher battery voltages. The end of the rapid drop which takes from 1.5 to 2.5 ns may be in the neighborhood of 40 to 60 volts depending on the switch. After this, the voltage tapers off at a rate

TABLE III — COMPARISON OF SWITCHES AT 278 VOLTS, MAX. CURRENT 1.86 AMPS. EFFECT OF DIFFERENT GASES WITH ONE CONTACT METAL AT ONE VOLTAGE

	Switch A	Switch C	Switch K	Switch E
Gas	97%N_2 + 3%H_2	Air	Helium	Hydrogen
Voltage of first glow (volts)	—	—	220*	242
Current in glow at end (amps)	—	—	0.39	0.24
Power in glow at end (watts)	—	—	85	58
Approx. duration of first glow (ns)	—	—	250	200
Voltage at end of second glow (volts)	—	—	206	—
Current at end of second glow (amps)	—	—	0.48	—
Power at end of second glow (watts)	—	—	99	—
Recovery of voltage to (volts)	—	—	262	278
Current at recovery (amps)	—	—	0.100	0
Power at end of recovery (watts)	—	—	26	0
Voltage at beginning of arc (volts)†	63‡	150§	137‖	102¶
Current in arc at start (amps)	1.43	0.85	0.94	1.17
Power in arc at beginning (watts)	90.5	128	129	120
Voltage at arc at 200 ns (volts)	24	20	120	56
Current in arc at 200 ns (amps)	1.70	1.72	1.05	1.47
Power in arc at end of 200 ns (watts)	41	34.5	126	82
Energy expended in 200 ns in nj	13,000	16,200	15,500	20,200
Voltage at beginning of alternate arc (volts) (B7 BL17)	—	62§	60‡	53‖
Current at beginning of alternate arc (amps)	—	1.44	1.45	1.50
Power at beginning of alternate arc (watts)	—	89	87	80
Voltage after 200 ns (volts)	—	10	20	15
Current after 200 ns (amps)	—	1.72	1.72	1.75
Power after 200 ns (watts)	—	34.5	34.5	26
Energy expended in 200 ns in nj	—	12,000	12,000	10,600

* Plate D, Nos. 129 and 432 — also compare with the MBV values of Table I.
† These are new results — compare with the usually accepted values of the order of 14 volts and 0.5 ampere for the short arc.
‡ Plate B, No. 594.
§ Plate C, No. 945.
‖ Plate D, No. 129.
¶ Plate D, No. 432.

which is approximately a straight line on the work diagram (CRO-2) ultimately ending in the 14-volt region before extinction, either by metallic closure, or by the arrival of the reflection from the end of L_1.

(5) The discharge may often persist for a time interval greater than $2T$. In this case the energy represented by the above straight sloping line in

this time interval may be calculated easily without integration. The height of the line represents the instantaneous power at this point so the average power is estimated by calculating the power in watts at the beginning of the straight line and again at the end of the 200 ns interval and multiplying the sum by 100. The result is the energy in nanojoules for the 200 ns time interval. These are the values given in Tables III and IV.

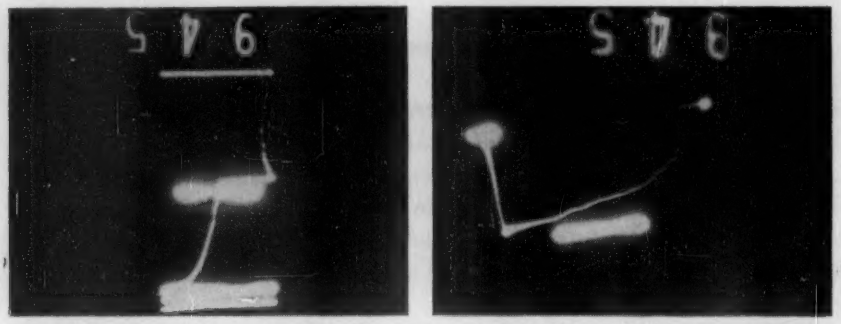
(6) Comparison of the photograph (Plate B, No. 594 is typical) shows the high initial voltage drop to take a time of about 2.5 ns which is a significant time interval and within the resolution of the scope. The energy associated with this process is about 270 nj. This represents the energy requires to initiate the arc process assuming the line gradually sloping from 60 to 14 volts represents an arc.

(7) At the highest voltage used (278) the greatest difference between the switches is revealed. The switch filled with hydrogen (Plate D, No. 432), reveals a pre-arc discharge at 242 volts which may be presumed an "abnormal glow". The transition to this voltage takes place at a much slower rate than the transition to an arc. Once formed it may persist or dwell for a considerable interval of time; often sufficient that interaction occurs between the glow and the voltages reflected from the end of the line. This may produce apparent inconsistencies between the patterns displayed by CRO-1 and CRO-2 as shown in Plate D, No. 432, and render interpretation of the patterns complicated.

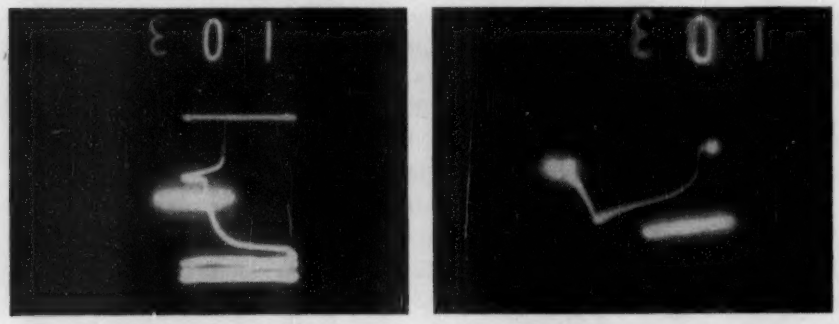
The helium switch yielded a two stage glow ending at 206 volts Plate D, No. 129, while the forming gas and air switches proceeded directly to the "arc" without a glow condition. The glow condition did not take place every operation but was present for about ½ the number of opera-

TABLE IV — BEHAVIOR OF ONE GAS (FORMING GAS) AND ONE METAL AT VARIOUS VOLTAGES

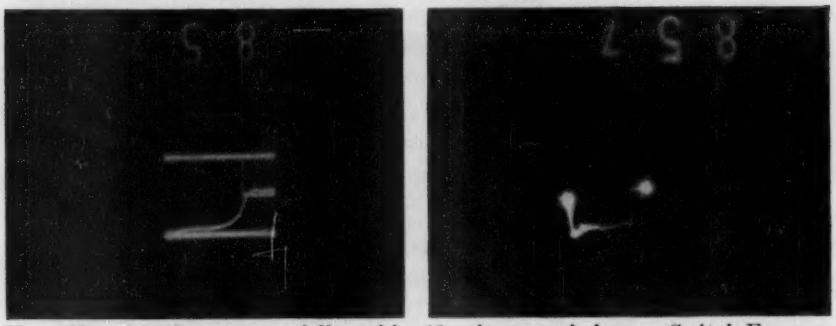
Behavior of One Switch (A)	Voltages			
	48	92	(190V)	(278V)
Max. closed cir. current (amps)	0.32	0.61	1.26	1.86
Voltage at end of sudden drop (volts)	14	14	35	63
Current in arc at same point (amps)	0.22	0.52	1.03	1.43
Power in arc at same point (watts)	3.2	7.3	36	90.5
Voltage at last stage of arc (volts)	14	14	14	14
Current at last stage of arc (amps)	0.22	0.52	1.17	1.76
Arc voltage at 200 ns (volts)	—	14	14	24
Arc current at 200 ns (amps)	—	0.52	1.18	1.70
Power in arc at 200 ns (watts)	—	7.3	16.5	40.5
Time required for sudden drop (ns)	—	1	1.5	2.5



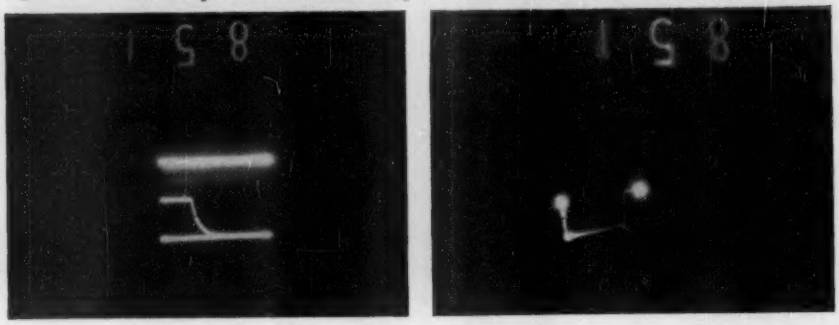
V_B = 278 volts — long arc — 20 volts 1.72 amps at 200 ns. Initial voltage 150 current initial .86 amperes. Note voltage plateaus at approximately 150, 104, 80 and 70 volts shown on both CRO-1 and CRO-2.
ΔL_2 = 5 ns — zero voltage base line superimposed for checking.



V_B = 190 — slow drop to 40V arc 0 — calibration line

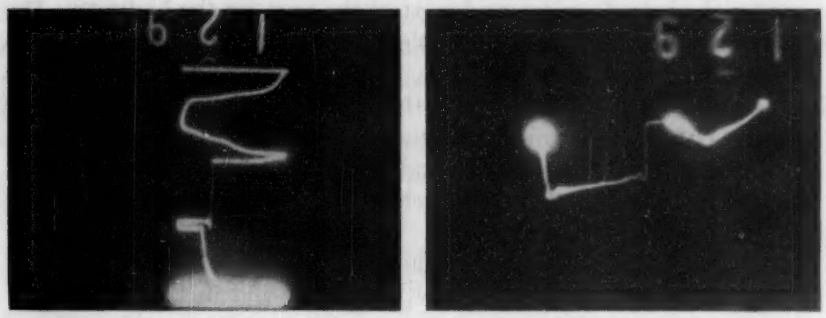


V_B = 92 — fast drop to zero followed by 13 volt arc and closure. Switch E.

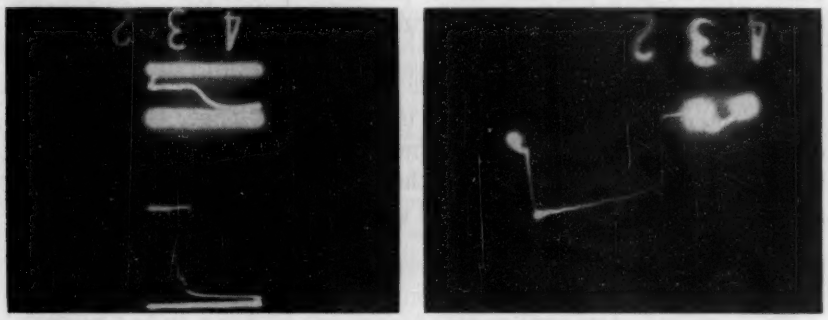


V_B = 92 — straight closure without arcing in 10^{-9} sec.

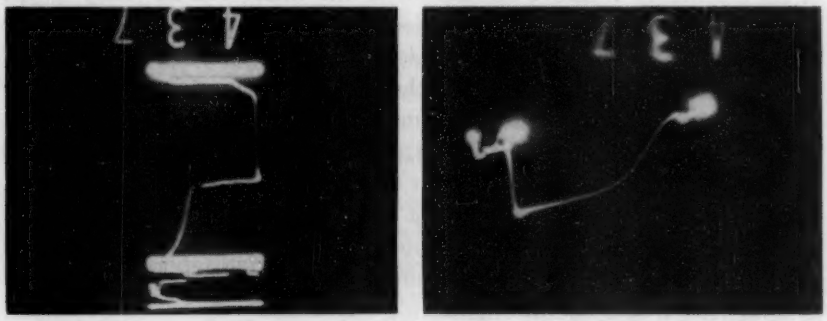
Plate C — Upper half, data on switch C (Air). Lower half, data on switch E (H_2).



V_B = 278. Slow glow type drop of 58 volts, dwell, further slow fall of 14 volts, dwell, slow recovery of 56 volts, dwell, rapid fall of 125 volts slow arc fall of 17 volts, eventual arc extinction at about 1000 ns. Typical pattern of this switch. Switch K.



V_B = 278 — Slow drop to glow at 242 volts — reflection of glow mixed with glow-dwell — fast drop of 176 volts to long arc final extinction. Switch E.



V_B = 278 — Two stage glow 255 and 248 volts — reflection coincident with second stage glow and reflection and ultra long arc with reflections. Switch E.

Plate D — Glow and arc discharges with reflection interference.

tions for the helium switch and most of the time for the hydrogen. The remainder of the operations resembled the discharge with forming gas.

(8) The air filled switch gave a fairly smooth and slow transition Plate C, Nos. 945 and 103, from the open condition to the usual arc voltages. Often the nature of the discharge changes from the slow type illustrated in Plate C, Nos. 945 and 103, to the fast type of Plate B, Nos. 594 and 698. Whether the switch is "activated" or not there is more than one discharge process.

(9) Another phenomenon of significance revealed by CRO-2 is the variable flow of charge in the glow discharge, and in the glow to arc transition. Sometimes the current is interrupted (no horizontal motion of CRO-2) followed by prebreakdown currents at roughly constant voltage (horizontal motion only) terminated by an arc. Examples are shown in Plate D, Nos. 129, 432, and 437. A possible explanation is that the discharge shifts from one surface irregularity to another but must start the breakdown process afresh.

These examples also illustrate the complication in the CRO patterns due to the discharges enduring for time intervals longer than $2T$ ns as described previously. It must be remembered CRO-1 has only two fixed references lines — the O line and the R line. The intermediate points are the resultant of the original and the reflected waves.

CONCLUSION

This circuit, and its oscilloscopic accessories, is shown to be a powerful tool for exploring contact phenomena both (1) under the ideal condition of an essentially resistive circuit where the interest is in study of the contact unmodified by frequency selective circuit elements and (2) under circuit conditions with frequency selective elements with reflections and other transmission phenomena. In the first condition, it is easy to relate the oscilloscope patterns to fundamental quantities. With the second condition interpretation of the patterns requires considerable detailed study. If the contact is not closed at the time the reflected wave arrives, the contact becomes the source for a new reflection and also may permit transmission in the forward direction. The study of such data should yield useful information as to the nature of the discharge process as well as information of a circuit engineering character.

The minimum of data presented here for illustration of the technique is the beginning of a collection which must be taken into account in any theories of contact behavior.

The author acknowledges his indebtedness to many individuals for help in the course of this development and in particular to S. O. Rice

and associates for assistance with the mathematical and transmission aspects of the problem. E. F. Thunder also gave valuable assistance in operation of the oscilloscope and made numerous improvements in its design.

APPENDIX A — DESCRIPTION OF OSCILLOSCOPE

Two Dumont K1068 oscilloscope tubes with metallized P11 screens were used in this work. The tubes are the post deflection acceleration type, with 4,000 volts (electron gun negative) used for the normal acceleration of the electron beam before entering the (ground Potential) deflection plate region plus 25,000 volts after leaving it. The deflection plates of these tubes give about equal beam deflections (Plate A) in either horizontal or vertical planes and the horizontal and vertical plates are shielded from each other. The connection to the deflection plates and shield are brought out directly through short leads to glove fastener terminals fused into the glass walls of the narrow neck of the tube.

The two tubes are mounted side by side about 6 inches between centers. Fig. 2 shows a schematic diagram of the relation between the tube face, the camera, and the counter.

The high voltages required by the post accelerator rings are supplied from a shielded potential divider with short connections to the terminals of the post deflection accelerator rings on the tube. Corona was suppressed by waxing over the connections from wires to tubes.

A separate shielded insulated grid to cathode battery for each CRO tube which could be varied in 1½-volt steps was found essential for stability.

The over-all design of the system enabled the two scopes to operate with satisfactory stability at the maximum beam intensity or whatever value was found desirable for the conditions at hand.

Sequence of Events in Oscilloscope and Relay Control Circuits (Fig. 1)

The closure of the hand operated control switch (or manual start switch) opens the camera shutters which remain open for a time interval (2–5 milliseconds) and then automatically close, after which time the counters step on to the next number. The same manual control contact operates a truly chatterless contact on a mercury contact relay* which in turn connects the operating winding of the relay under test to a local battery. The closure of the mercury contact also causes a flip-flop circuit

* Brown, J. T. L., and Pollard, C. E., Recent Developments in Relays — Mercury Contact Relays, Elec. Eng., pp. 1106–1109, Nov. 17, 1947.

to start measuring off an appropriate delay interval, (operate delay time of the test switch is about 1 millisecond) after which a second flip-flop circuit turns on the beam current of the oscilloscope, then turns it off a few microseconds later to prevent screen damage. The delay of the delay flip-flop* is adjustable only in steps so that a continuously variable delay network is also necessary so that the contact of the test relay may be adjusted to close during the time interval the oscilloscope beam is on. This arrangement is used so that no load is placed on the contact other than the test circuit in which it operates. The CRO beam on period and the test contact transition interval are thus independent but arranged to coincide. If the test contact exhibits multiple closures (chatter) the delay networks can be adjusted so that any one of them can be observed if desired. In this case the first closure only is considered.

APPENDIX B — THE PHOTOGRAPHIC RECORD

The arrangement of tubes and camera and counters is shown schematically in Fig. 2. The cameras are Kodak 35 bodies fitted with special lenses and operated by magnetic solenoids. The counters are provided with a viewing prism and their own illuminating lamps. In photographing the flying spot on the P-11 screen a 1/1 image/object ratio was used, in order to obtain a large photographic record. Also, it was found necessary that the image of the flying spot be in very sharp focus on the film to avoid gaps in the record of high velocity beams due to the lack of contrast between the illuminated region and the adjacent regions.

Two 50-mm $f/2$ Raptar lenses were mounted front-to-front as shown in Fig. 2, giving a lens corrected for focusing divergent light (from the beam spot) instead of *parallel* light as is usual with camera lenses. The arrangement also gives an increased ratio of diameter to focal length as well as sharper focus.

The film was Eastman Kodak Linograph Pan film LP-135, developed with a special developer D19A† which increases the effective speed of the film in the fast or faint parts of the trace.

The numbers on the film are upside down for mechanical reasons.

APPENDIX C — CIRCUITRY AND SOURCES OF ERROR

Circuit Restrictions and Modifications

For the short time intervals studied here the inductance and capacitances of very short lengths of wire (even one cm) are important. Most of the inductive reaction of resistances and lead wires can be compensated

* Term used here refers to a one shot multivibrator circuit which is arranged to turn a current on and then turn it off after a preset time interval.

† Methods of Increasing Film Speed, J. Photographic Soc. Am., **12**, pp. 586–610.

for by making them the central conductor of a coaxial line structure by providing a coaxial sheath with the proper diameter.

In the oscilloscope itself, however, one has a parasitic circuit comprising the capacitance of the glove fastener connections passing through the wall of the glass — the inductance of the lead wires inside the glass to the deflection plates, and the capacitance of the plates themselves. This circuit is shown in Figure 7(b) and is mechanically and electrically unchangeable with available CRO tubes. This circuit has very little damping, thus if a very sharp pulse is applied to the scope plates this circuit tends to ring and superimpose its own oscillations which interfere with the interpretation of the record. At best they can be used as a reference time scale once their frequency has been determined. Connecting too high a resistance R_0 outside this circuit only slows down the rate of discharge of the total capacity of the combination and damages the high frequency response of the over-all system.

Capacitances also have residual inductance due to the length of lead wires ordinarily required to connect them in the circuit. There are, however, "feed-through" coaxial type condensers which practically eliminate the effects of lead inductance.

With these facts in mind the oscilloscope plate input transformer circuit details must be considered. The usual simple diagram, shown in Fig. 7(a), must be replaced by the circuit shown in Fig. 7(b) for the vertical plates of either CRO-1 or CRO-2. The initial corresponding horizontal deflection plate circuit of CRO-2 is shown in Fig. 9(a). Because of the above considerations, it was modified along the lines indicated in Fig. 9(b), providing the necessary cavity to terminate the line L_1 in a resistance and capacity only. This reconstruction is not deemed necessary for the vertical plates as the reactance of $(R_0 + X_L)$ could be matched to $C_0 = C_f + C_d$ and the over-all Q of this circuit made nearly equal to unity so oscillations are not excessive. On the other hand, the parasitic reactances of Fig. 9(a) produce definitely reproducible effects like those shown in Plate B, No. 371, rendering the work diagram uninterpretable. A typical effect obtained with the reconstructed circuit, Fig. 9(b), are shown in Plate B, No. 594. It will be seen that the circuit simplification and compensation has eliminated all the difficulties from parasitics except those due to the internal structure of the vertical plates of the oscilloscope. The resonant frequency (about 400 megacycles) of this parasitic circuit is determined in terms of ΔL_2.

Sidestepping the Starting Spot Halo

In the balanced condition ($L_1 = L_2$) a portion of the initial record of the motion of the CRO-2 beam spot is rendered unobservable by the large

halo caused by the beam resting at the starting position for a relatively long time. This undesirable effect may be sidestepped by experimentally choosing an appropriate and small value of ΔL_2 say corresponding to 5 ns (Plate D, No. 594) and leaving it in the circuit. In this case the initial motion of the beam starts horizontally from the starting dot and comes out of the halo for a short distance ΔL_2 when it proceeds to fall as before. The corresponding times are measured from this instant up to the time $2T$.

Comment on the CRO deflection Plate System — Error from Finite Beam Velocity

With the 4,000-volt first acceleration voltage, the velocity of the beam electrons through the deflection plate region is 3.78×10^9 cm/sec giving a transit time of the electrons between the plates of about 0.4 ns which is the minimum rise time which can be observed with these tubes.

Errors in Energy Measurement By CRO-2

(1) In the work diagram, it has been assumed that the voltage drop in the COLS contact traces out an area corresponding to a significant amount of energy. In the case of an ideal or instantaneous closing contact, there would still be some small area traversed even if there were no energy loss in the COLS. This area is the area under the exponential decay line drawn through the starting dot and corresponding to the discharge of the vertical plates of CRO-2 alone from the battery voltage $V_B/2$. The area under this line corresponds to 4 nanojoules (nj)* for a 45 volt V_B and about 160 nj for the 280-volt battery. Energies in excess of these values measured with this system may be ascribed to the dissipation of the contact.

(2) For sufficiently large values of R_2C_1 (Fig. 4) with $R_2 = Z_0$ the *charge* is accurately measured by the change in voltage of C_1 (neglecting C_0). However, for small values of R_2C_1 the change in voltage of C_1 becomes an appreciable fraction of and opposing the voltage producing i.

TABLE C-I

n	Q_{C1}/Q_∞	n	Q_{C1}/Q_∞
1/10	0.976	2	0.633
1/2	0.885	3	0.518
1	0.786	4	0.382

* 100 *nj* equal 1 erg.

Hence i is decreased and the change in voltage is a measure of the reduced i. Thus with small capacities the area traced by CRO-2 is always less than the true area were a large enough capacitor used. In general the correction factor cannot be calculated. However, for a *constant unit* contact voltage step the charge Q_{C1} is

$$2i_0R_2C_1(1 - e^{-(t-T)/2R_2C_1})$$

(which is restricted to values of time between $t = T$ and $t = 3T$, and i_0 is the initial current). With an infinite capacity the charge is

$$Q(C = \infty) = i_0(t - T)$$

The ratio of

$$(Q_{C1}/QC = \infty) = \left(\frac{2R_2C_1}{(t - T)}\right)(1 - e^{-(t-T)/2R_2C_1})$$

If

$$\frac{t - T}{R_2C_1} = n$$

then

$$(Q_{C1}/Q\infty) = \frac{2}{n}(1 - e^{-(n/2)})$$

Values of the ratio as a function of n are given in Table C-I. The error is $1 - Q_{C1}/Q^{\infty}$ and is always in the direction of measuring less charge than actually passed through the contact.

To obtain high time resolution in the early parts of the discharge a small value of R_2C_1 should be used. In this data R_2C_1 is 7.8 ns. The energy dissipated is always at least as great $(Q_{C1})/(Q^{\infty})$ as that indicated by the area. For the later and slower parts of the discharge appropriately large values of R_2C_1 are used.

(3) Within the time limits from zero to $2T$, where $T = 100$ ns the resolution is excellent. However, one cannot stretch T out to long time intervals (of the order of several microseconds) without attenuation in the cable rounding off of the corners of the voltage step functions.

Dual Voltage Operation of Relays and Crossbar Switches

By. A. C. MEHRING and E. L. ERWIN

(Manuscript received June 27, 1955)

The operating speed of a relay or switch varies with the amount of electrical power supplied to it. For a given device, the speed can be increased by raising the voltage. Since the electrical energy is converted to heat, the speed is usually limited by the heat dissipating characteristics of the device.

A means for operating relays and switches on high voltage and then changing the circuit to hold them operated on low voltage is described. The circuit is switched by means of a solid state junction diode.

Reduction of relay operate times to half their former values, and reduction of switch operating times to one-third their former values can be obtained with commercial circuits. The wire-spring markers of the No. 5 crossbar system are equipped with dual voltage circuits for speeding the operation of the crossbar switches of the switching network.

I. INTRODUCTION

Making a connection from a customer's telephone to any of the other fifty million telephones in the United States requires a large and complex data processing or switching system. The equipment for this switching system is located in thousands of telephone central offices throughout the country. The equipment in a modern telephone central office may be listed within one of two catagories, the equipment which provides the "talking" or transmission connections through the office, and the control equipment which selects and controls these connections for the talking transmission facilities. Telephone switching offices in which the control equipment has been centralized are labeled "common" control switching offices since the centralized control equipment is available and used "in common" by all of the customers of the office.

In these common control switching offices information which is received from the calling customer is stored in "memory-mechanisms". Other memory devices store information about the switching network which is

needed for routing the call to the called customer. A few common control circuits interpret the stored information, test for idle connecting paths and operate the switching mechanisms needed to establish a transmission interconnection from the calling customer's telephone to the called customer's telephone. Where the called customer is associated with another telephone central office, a transmission connection to this distant office is selected and information is transmitted through this transmission connection so that the distant office may complete the connection to the called customer.

The number of operations required of the common control equipment is determined by the total number of interconnections for which this common control equipment must provide the control facilities. Therefore, the number of common control units which will be required will increase or decrease as the rate of interconnections is increased or decreased. However, the number of common control units required will depend upon the speed of operation of each of these common control units. The greater the speed of operation, the smaller the number of common control units which are necessary. If this common control equipment can be made to operate fast enough, a single control unit would be sufficient for an entire switching office. Therefore, there is a strong economic incentive to develop fast operating control circuits.

One of the most complex and costly control circuits in a modern switching office, such as the No. 5 crossbar office, is designated a marker. A typical No. 5 crossbar office would require 6 markers or 20 frames (23 inch) of marker equipment.

Although this marker guides hundreds of separate actions it is fast in its operation, requiring less than half of a second to serve each customer's call. If the marker operation could be made faster, fewer markers would be required. It has been estimated that if its operation could be speeded by a factor of 10 then only one marker could handle the traffic for a large office. Therefore, the cost of a switching office is not only dependent upon the manufacturing and maintenance costs of relays and crossbar switches but also upon their operating speed.

To build fast operating markers, the basic building blocks of the switching office must also be fast operating. In a No. 5 crossbar office two basic building blocks are the electromagnetic relay and the electromagnetic crossbar switch which are shown on Figs. 1 and 2. A typical office would contain about forty thousand relays and a thousand crossbar switches. Four million relays and a hundred thousand switches are capable of serving about one million customers using No. 5 crossbar offices.

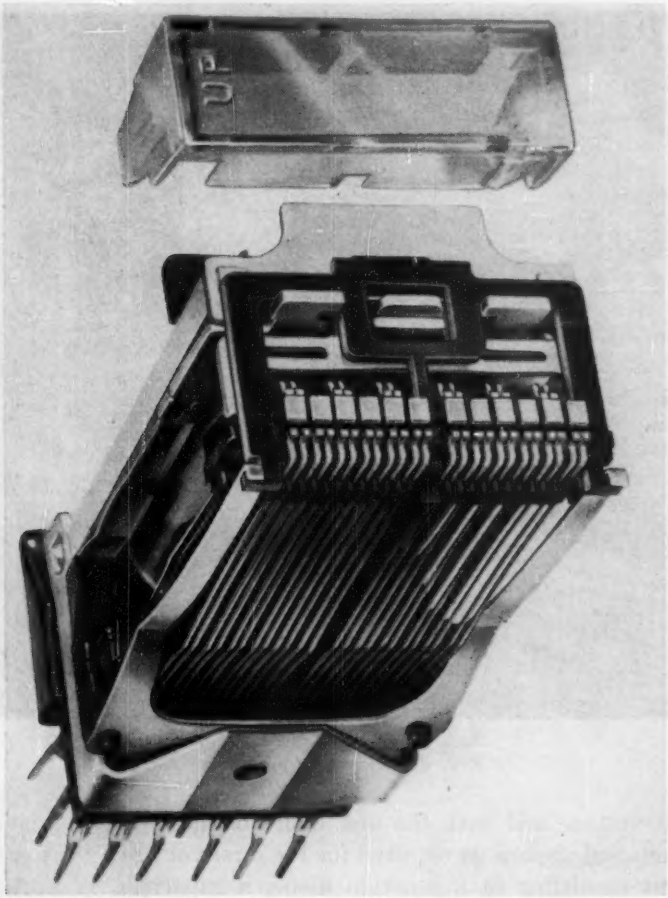


FIG. 1 — Modern telephone type relay known as type AJ or AF.

To achieve faster operating relays and crossbar switches a new dual voltage circuit has recently been developed. With the dual voltage circuit, a high voltage will be momentarily applied to the magnet of a relay or crossbar switch, producing a very fast operating device. The high voltage, which is obtained from a capacitor which has been precharged to this high voltage, will be applied during the operation of the relay device. A short time after the crossbar switch or relay operates, a germanium junction diode will automatically switch from the high-voltage capacitor circuit to the standard voltage (48 volts) for holding operated the device. Crossbar switch or relay operating circuits with the usual

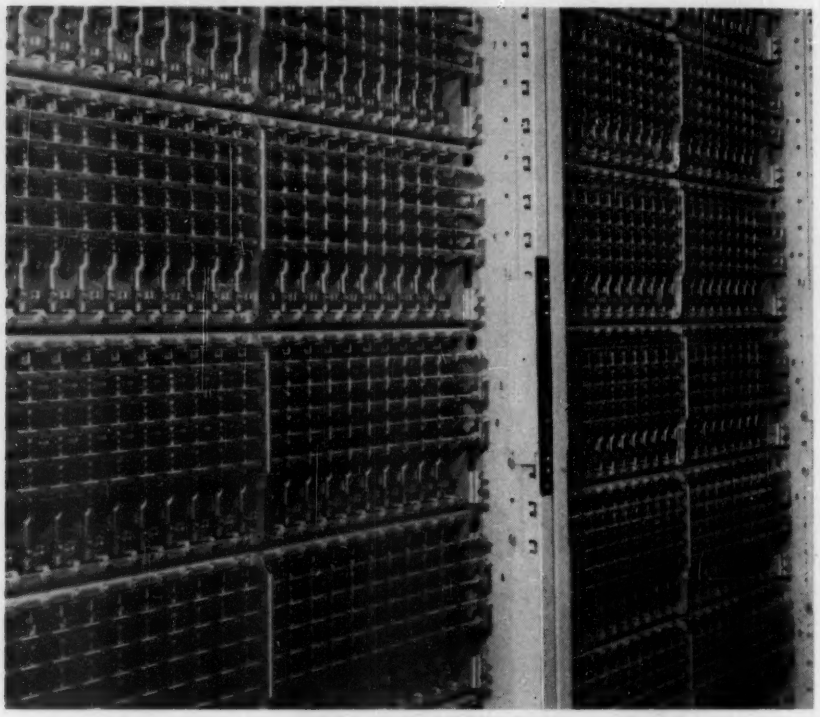


FIG. 2 — Crossbar switches.

constant voltage and with the new dual voltage are shown in Fig. 3. The additional apparatus required for the dual voltage circuit is a compact unit consisting of a junction diode, a capacitor (of about 4-mf capacity) and a resistor as shown in Fig. 4.

Without the new dual voltage circuit, the operating time of a typical crossbar switch hold magnet is 0.055 second or for a typical wire-spring relay, it is 0.0055 second. With this dual voltage circuit the crossbar switch hold magnet will operate in 0.018 second, which is about one-third of its former operate time, and the relay will operate in 0.0029 second, which is about one-half of its former operating time.

The dual voltage circuit is now in use for fast operation of crossbar switch hold magnets of No. 5 crossbar offices. Fewer markers are necessary with the faster operating crossbar switches and faster markers. The dual voltage circuit causes a small increase in the cost of markers but this is insignificant compared to the saving resulting from fewer markers. It is expected that the future application of the dual voltage circuit to

relay circuits and to other types of switching offices will produce additional savings. Table I shows the gain in speed which can be obtained with typical telephone relays.

II. FUNDAMENTALS OF SWITCHING RELAY OPERATION

It is a well known fact that the operating time of a crossbar switch or relay will decrease, as the voltage applied to the magnet winding is increased or as the resistance of the magnet winding is decreased. For en-

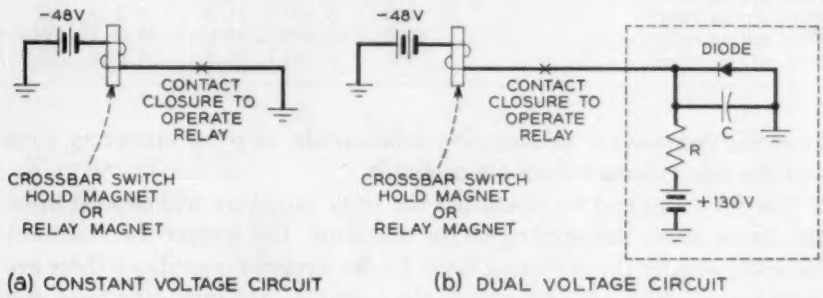


FIG. 3 — (a) Constant voltage circuit. (b) Dual voltage circuit.

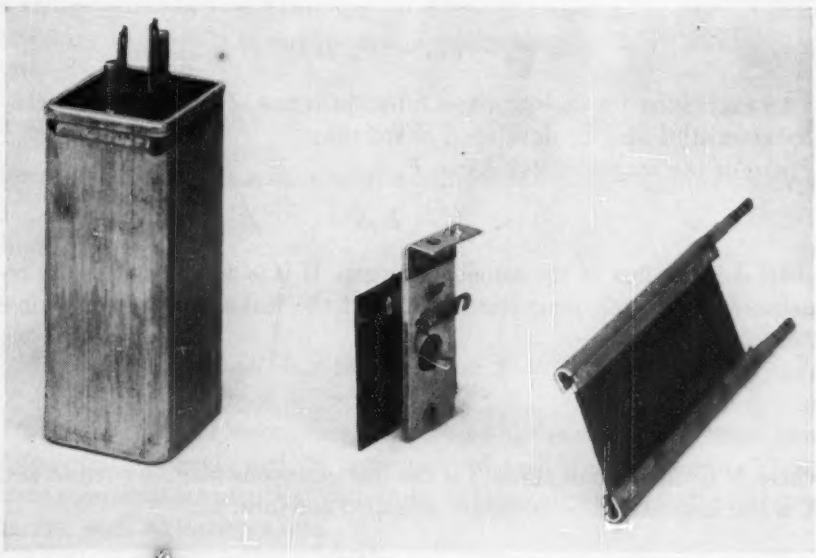


FIG. 4 — The apparatus of a dual voltage circuit.

TABLE I — COMPARISON OF OPERATING SPEED OF TYPICAL RELAY DEVICES

Type	Relay Device		Operate Time Milliseconds		Per Cent Operate Time Saving
	Magnet Resistance in Ohms	Magnet Turns	With Constant Voltage Circuit	With Dual Voltage Circuit	
Crossbar switch hold magnet	1250	14,000	55.0	18.0	67
Wire spring relay	16	1,580	3.3	1.7	49
Wire spring relay	270	2,110	5.5	2.9	47
Wire spring relay	700	5,050	11.0	4.2	62
Wire spring relay	2500	19,400	15.0	6.0	60
Flat spring relay	400	5,300	10.6	4.6	57

gineering purposes a quantitative relationship of relay operating time and the relay characteristics is desirable.

The time required for operating the relay armature will depend upon the forces which are applied to the armature, the greater the resultant force the smaller the operating time. Under dynamic conditions there are three forces which are acting upon the armature, the operating force due to the magnetic flux F_ϕ, the restraining spring forces, F_{RS}, and the armature mass inertia force, F_M. The resultant of these forces must be zero, therefore:

$$F_\phi - F_{RS} - F_M = 0$$

An expression for each of these forces in terms of the circuit or relay characteristics may be developed as follows:

(a) For the magnetic flux force, F_ϕ:

$$F_\phi = K_1\phi^2$$

where ϕ is the flux of the armature airgap. If it is assumed that the reluctance of the iron magnetic circuit and the leakage flux is negligible then:

$$\phi = \frac{K_2 N i}{X}$$

where N is the magnet turns; i is the instantaneous magnet current and X is the instantaneous armature airgap. Therefore:

$$F_\phi = K_3 \left(\frac{Ni}{X}\right)^2$$

(b) For the restraining spring force, F_{RS} :

For a typical relay it may be assumed that the restraining spring force increases directly as armature airgap decreases or:

$$F_{RS} = \frac{K_4}{X} + K_5$$

(c) For the armature mass inertia force, F_M :

$$F_M = M \frac{d^2X}{dt^2}$$

where M is the armature mass.

Therefore the equation:

$$F_\phi - F_{RS} - F_M = 0$$

May be written as:

$$K_3 \left(\frac{Ni}{X}\right)^2 - \frac{K_4}{X} - K_5 - M \frac{d^2X}{dt^2} = 0 \tag{1}$$

This is known as the mechanical differential force equation of a relay. This equation by itself cannot provide a relationship between the instantaneous armature airgap X and time t since the electrical current i is an unknown factor. Therefore, to define the electrical current i it is necessary to develop an electrical differential equation as follows:

$$iR + N \frac{d\phi}{dt} - E = 0$$

where R is the circuit resistance and E the continuously applied voltage.

Since $\phi = \dfrac{K_2 N i}{X}$

$$iR + K_2 N^2 \frac{d\left(\frac{i}{x}\right)}{dt} - E = 0 \tag{2}$$

The dynamics of electromagnetic relays are governed by these two differential forces (1) and (2) one electrical and one mechanical. These basic equations are identical with those of other electromechanical transducers, such as loudspeakers.

An approximate solution of the two differential equations which is suitable for some engineering purposes for electromagnetic devices has

been provided by R. L. Peek, Jr.* It defines the operating time as follows:

$$t\,(\text{time}) = K_6\left(\frac{E^2}{R}\right)^{-1/3} + K_7\left(\frac{E^2}{R}\right)^{-1} \tag{3}$$

The constants K_6 and K_7 are determined by the mechanical, magnetic and electrical structure of the relay device. Relay devices with different mechanical, magnetic or electrical structure will be defined by equations with different constants of K_6 and K_7. Two relays with identical mechanical and magnetic structure, but with different magnet coils, may be defined by the same constants K_6 and K_7 if the coil constant, which is N^2/R, is the same for both relays.

A typical crossbar switch hold magnet which consists of 14,000 turns with 1,250 ohms resistance operates in 0.055 second with 48 volts applied and operates in 0.014 second with 178 volts applied from a constant source. The actual experimental operating times, as a function of E^2/R, are shown in Fig. 5. These experimental data closely approximate the equation:

$$t = 0.03\left(\frac{E^2}{R}\right)^{-1/3} + 0.05\left(\frac{E^2}{R}\right)^{-1}$$

where t is expressed in seconds. For any type crossbar switch hold magnet for which the coil constant, N^2/R, is $14{,}000^2/1{,}250$, the operating time

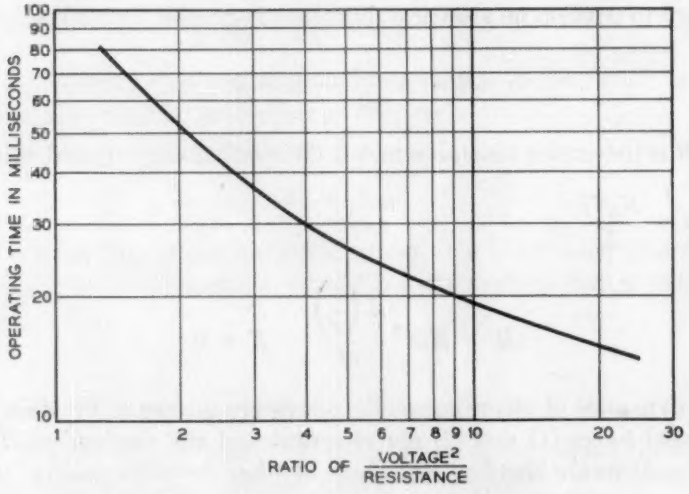


FIG. 5 — Crossbar switch hold magnet operating time.

* R. L. Peek, Jr., Estimate and Control of the Operate Time of Relays, B.S. T.J., **33**, Jan., 1954.

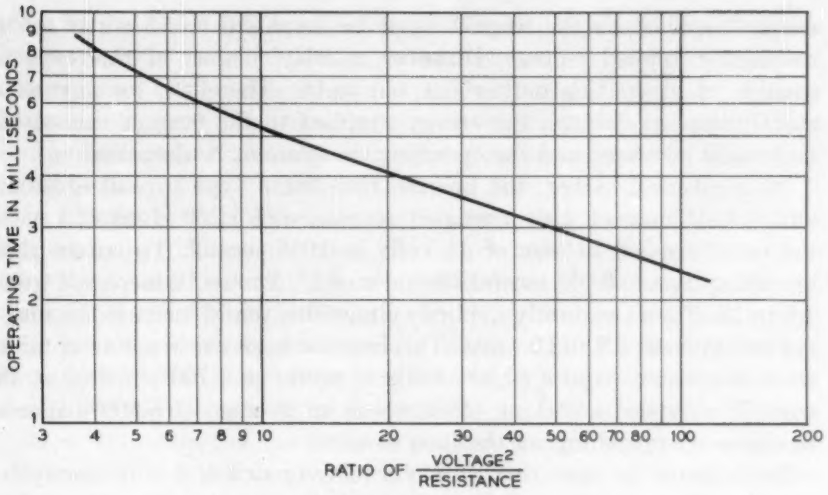


FIG. 6 — Relay operating time.

for any applied voltage and for any coil resistance may be evaluated from Fig. 5.

A typical relay of the wire-spring type which consists of 2,110 turns with 270 ohms resistance operates in 0.0055 second with 48 volts applied and operates in 0.0023 second with 178 volts applied from a constant source. The actual experimental operating times, as a function of E^2/R, are shown in Fig. 6. These experimental data very closely approximate the equation:

$$t = 0.01 \left(\frac{E^2}{R}\right)^{-1/3} + 0.005 \left(\frac{E^2}{R}\right)^{-1}$$

where t is expressed in seconds.

For any wire-spring type relay (with intermediate travel) for which the coil constant, N^2/R, is 15, the operating time for any applied voltage and any coil resistance may be evaluated from Fig. 6.

III. THE DUAL VOLTAGE CIRCUIT

The operate time is 0.0055 second for the wire-spring type relay with a coil resistance of 270 ohms when the usual voltage of 48 volts is applied. If it is desired to reduce this operate time to 0.0029 second, as an example, the ratio of E^2/R must be increased from 8.5 to 55. For a constantly applied voltage the ratio of E^2/R has the dimensions of power in watts. Therefore, to decrease the operating time to 0.0029 second, the

energy supplied to the magnet must be increased to 55 watts for a constantly applied voltage. However, a relay magnet of this type is capable of dissipating only about ten watts. Therefore, for any constantly applied voltage, the energy supplied to the magnet cannot be materially increased and the operating time cannot be decreased.

As mentioned earlier, the operate time for a type typical crossbar switch hold magnet with a magnet resistance of 1,250 ohms and with the usual applied voltage of 48 volts is 0.055 second. To reduce this operating time to 0.018 second, the ratio of E^2/R must be increased from 1.8 to 12. For a constantly applied voltage this would increase the magnet energy from 1.8 to 12 watts. This increase in power is not acceptable since this would require 80,000 watts of power or 1,700 amperes at 48 volts in a typical switching office where an average of 6,800 magnets would be energized during the busy hour.

To decrease the operating time of a relay or switch it is necessary to abandon the restrictive concept of circuits with only one voltage. Instead, the basic requirements of a fast operating device dictate that two different or dual voltages be used for energizing it. Several years ago J. C. Rile suggested that a circuit of this type could be used to improve the operating times of switches in crossbar systems.

With dual voltage operation a voltage of large magnitude is momentarily applied to the magnet. The magnitude of this momentarily applied voltage will quickly decay and will be supplanted by a lesser voltage which will be constantly applied to the magnet to hold the relay or switch operated until it is to be released. In this way, the operating speed will be determined by the magnitude of the larger voltage which is momentarily applied to the magnet. Because this larger voltage is applied for a very short period of time the magnet heating which it produces will be negligible. Therefore, the magnitude of this larger voltage is not limited by magnet heating. The heating power dissipated in the magnet will be primarily the result of the lesser voltage when constantly applied to the magnet. This lesser voltage can be small and need be sufficient only for holding the relay or switch operated. For dual voltage operation, it would be possible to develop a circuit using relay contacts to switch from the larger voltage to a lower voltage after the crossbar switch or relay had operated. However, this circuit would introduce two problems; first, the switching of a high voltage and current with relay contacts, and second, the fire hazard from overheating the magnet of the crossbar switch or relay if the switching contacts failed to promptly remove the larger voltage. To overcome these problems a dual voltage circuit as shown in Fig. 3(b) was developed which operates the crossbar switch or

relay by a momentarily applied voltage obtained from a capacitor which has been precharged to the larger voltage. Since the capacitor has limited energy, there is no possibility of overheating the relay. No relay contacts are required to switch from the initially applied voltage to the lesser continuously applied voltage since the germanium junction diode performs this switching function.

In a typical circuit with a constant voltage, the closure of the controlling contact or contacts applied ground potential to the magnet of the relay device and thus 48 volts are applied. Because of the inductance of the magnet the current build-up is slow. This slow current build-up results in a slow operating relay device (as compared with the dual voltage circuit). With the dual voltage circuit for high speed operation, the relay device will be operated by a momentarily applied voltage of 178 volts (130 + 48) instead of the usual 48 volts. After operating, the relay device will be held operated with 48 volts. The voltage of 178 volts which is momentarily applied to the relay device is obtained from a capacitor which is precharged. As shown in the dual voltage circuit in Fig. 3(b),

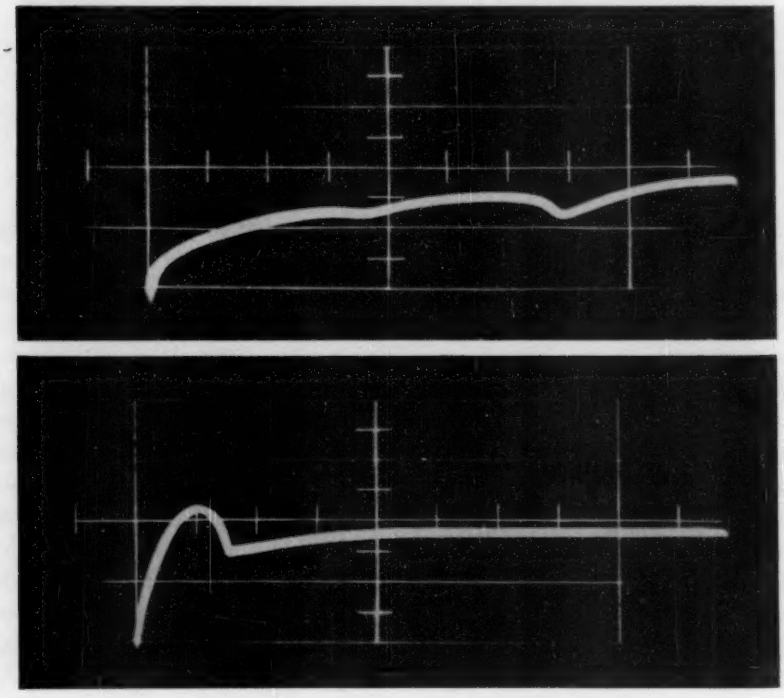


FIG. 7 — Current in crossbar switch hold magnet. Time division is 0.01 sec.

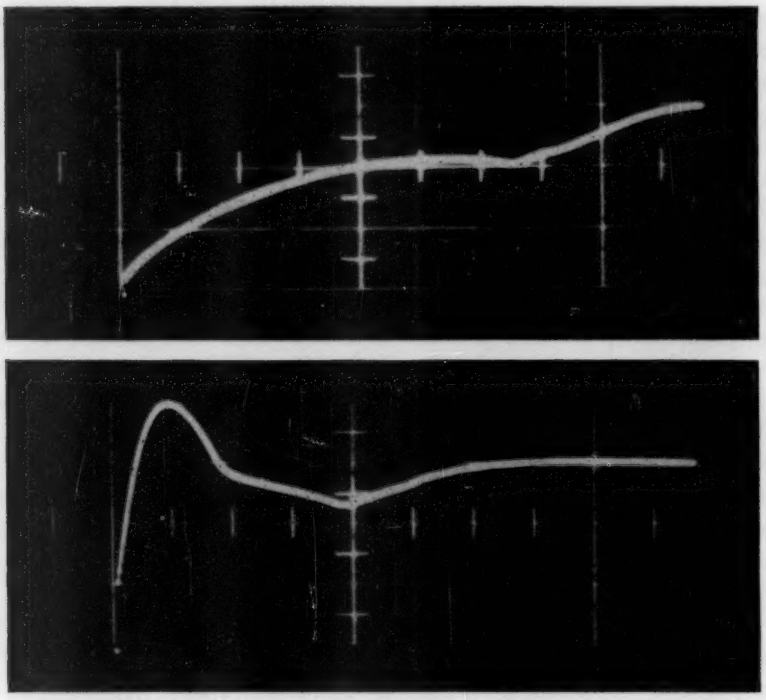


FIG. 8 — Current in typical relay magnet. Time division is 0.001 sec.

the +130 volt potential precharges the capacitor to +130 volts before the relay is to be operated. The closure of the controlling contact or contacts applies this +130 volts to the magnet winding already connected to −48V and thus a 178 volt potential is momentarily applied. Because of the larger voltage which is applied to the magnet winding the current will build-up rapidly resulting in very fast operation. This rapid build-up of current for the dual voltage circuit is contrasted with the usual slower build-up of current as shown in Fig. 7 for the crossbar switch magnet and in Fig. 8 for the wire-spring type relay magnet. The voltage which is applied to the magnet winding will have an initial magnitude of +130 volts but will decay rapidly to ground potential, as shown in Fig. 9 for the crossbar switch magnet and in Fig. 10 for the relay magnet. While the capacitor is discharging its energy into the relay, the voltage across the capacitor will drop from +130 volts toward a −48 volt potential. As long as the capacitor voltage is of a positive polarity the diode will act as an open switch contact and no current will flow through the diode. However, when the capacitor voltage has decreased to about −0.5 volt

the diode will act as a closed switch contact and current will flow through the diode. The current which flows through the diode will hold the relay or crossbar switch magnet operated.

If the capacitance in the dual voltage circuit is very large, the voltage applied to the magnet winding would be 178 volts and would decay very slowly to 48 volts. With this very large capacitance, the operating time of the crossbar switch or relay would very closely approach the operating time which is obtained when a continuous voltage of 178 volts is applied. However, for smaller and more practical values of capacitance the operating time would be somewhat greater than for the continuously applied voltage of 178 volts. Fig. 11 shows a graph which correlates for

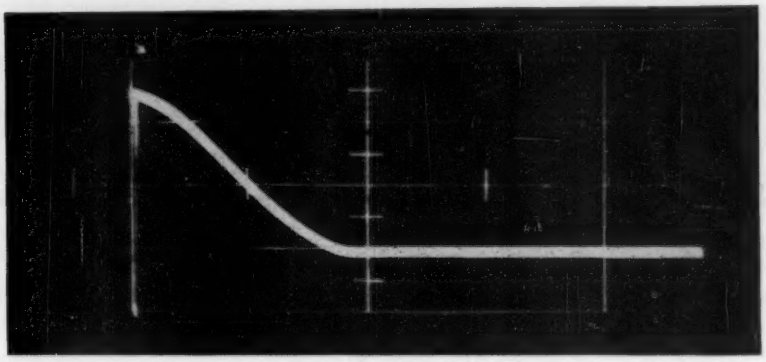


FIG. 9 — Voltage across crossbar switch magnet with dual voltage circuit. Time division is 0.01 sec.

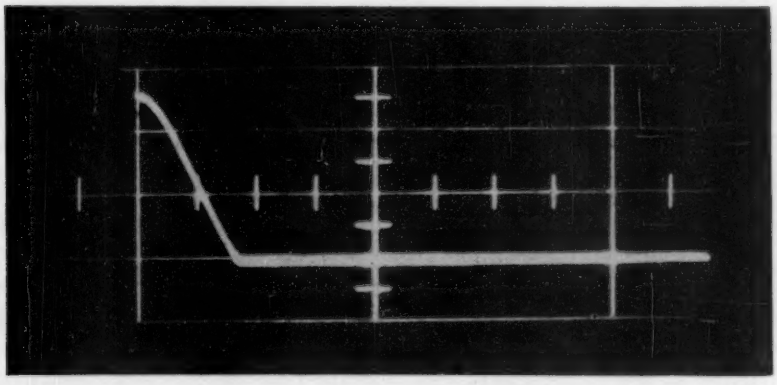


FIG. 10 — Voltage across relay magnet with dual voltage circuit. Time division 0.001 sec.

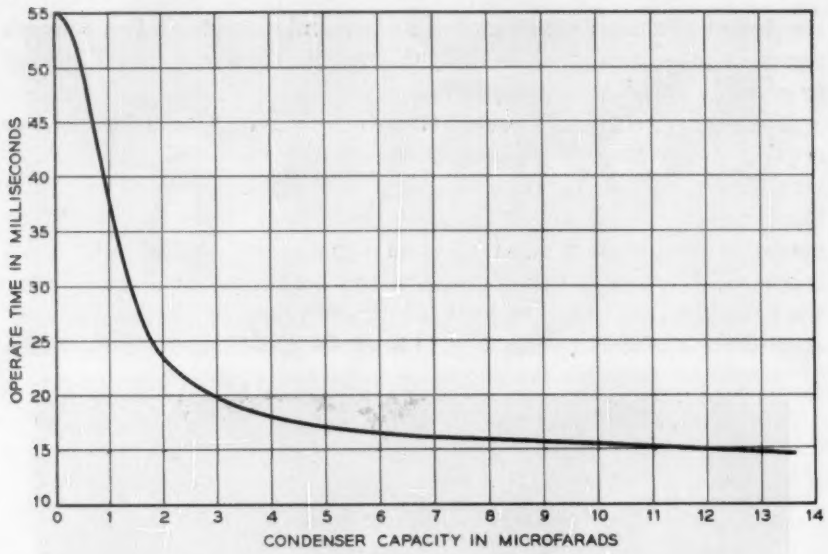


FIG. 11 — Operate time of crossbar switch as a function of condenser capacity.

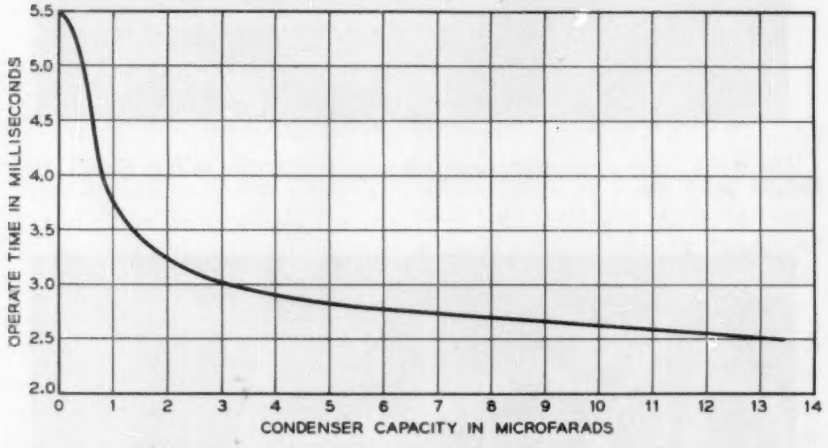


FIG. 12 — Operate time of typical relay as a function of condenser capacity.

various values of capacitance the corresponding operate time of the crossbar switch hold magnet while on Fig. 12 is shown a similar graph for the wire-spring type relay. In a typical dual voltage circuit design with a capacitor of 4.0 microfarads, the operate time of the crossbar switch magnet is 0.018 second while the operate time of the relay is 0.0029 second.

IV. APPLICATION TO NO. 5 CROSSBAR MARKERS

A typical No. 5 crossbar office may contain 1,000 crossbar switches with 20,000 hold magnets. To speed up the operation of these magnets through the use of the dual voltage circuit, it is *not* necessary to provide a capacitor-diode network for each magnet. The network is part of the marker. The magnets have the −48 volt central office battery connected to one terminal of their windings. The other winding terminals are wired to the contacts of connector relays so that these circuits can be extended into the markers. When a marker establishes a connection, it operates the proper connectors, through which it reaches the magnets it wishes to operate. Without the dual voltage feature, the marker would place ground on the leads corresponding to magnets to be operated. For dual voltage operation the marker connects the capacitor-diode network, instead of ground, to the magnet lead. Having the dual voltage network in the marker permits the network to be successively used for the operation of many magnets. The time required to recharge the capacitor for its next use is not a serious problem here. There is adequate time to recharge the capacitor while the marker is performing other functions. The charging time is determined by the size of the resistor and capacitor.

The switching network of a No. 5 office (Fig. 13) is a three-stage arrangement. A connection requires a line link, a junctor and a trunk link. The combination of these three elements is called a channel. Before operating any magnets, a marker tests the elements of the available

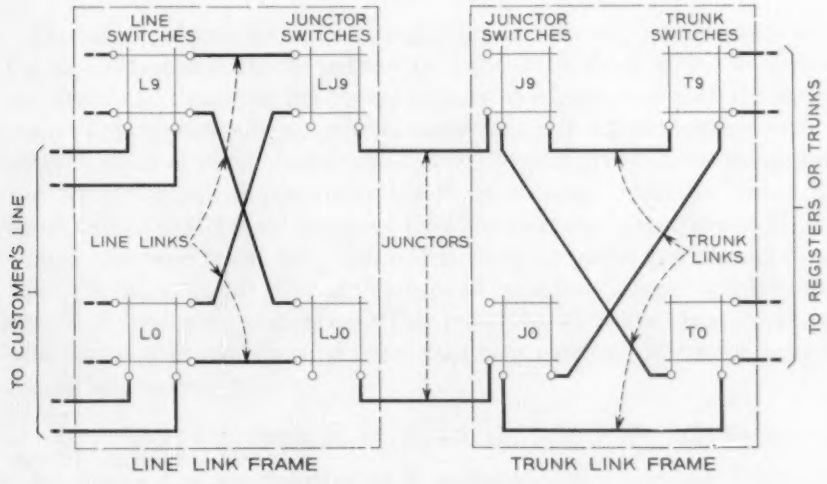


FIG. 13 — No. 5 crossbar switching network.

channels and selects the channel to be used. Having made channel selection and operated the proper select magnets, the marker then operates all the appropriate hold magnets simultaneously. Three capacitor-diode circuits are required in each marker to accomplish this simultaneous operation. The switch crosspoints are closed by the operation of the hold magnets. Closure of the crosspoints connects all the hold magnets of the channel together in multiple so they can be held by ground supplied from the register or trunk circuit after the marker disconnects.

ACKNOWLEDGMENTS

The authors wish to thank J. A. Ceonzo for doing the laboratory work on which the commercial development was based, and N. P. Santoro for obtaining the laboratory data used in this text.

Digital Memory in Barrier-Grid Storage Tubes

By M. E. HINES, M. CHRUNEY, and J. A. McCARTHY

(Manuscript received July 2, 1955)

In barrier-grid storage tubes, a cathode ray beam is used to deposit electrostatic charge on a dielectric surface. The same beam is utilized to detect and remove the charge at a later time. This paper describes the mode of operation of such tubes when used for storage of binary digital information. The sources of variation in the output signal are discussed. A basis is given for determining the probability of error due to amplifier noise. The storage capacity and program of operation are determined by the requirement of error-free performance. A discussion is included of data obtained from experimental tubes capable of storing 16,000 bits of information with reading or writing times of about one microsecond. A new type of target-reading circuit which is particularly suited to binary storage is also described.

I. INTRODUCTION

Storage of information in barrier-grid tubes is accomplished by depositing an electrostatic charge pattern on a dielectric sheet with a cathode-ray beam and detecting the charge pattern at a later time with the same beam. This type of tube obtains its name from a fine grid placed immediately in front of the dielectric sheet. The "barrier-grid" serves to inhibit the redistribution of secondary electrons emitted from the dielectric sheet and to shield the drift space from electrostatic disturbances which would otherwise result from the stored charge pattern and the application of writing signals. Use of this type of tube for "image" storage was described by Jensen and others.[1] This paper describes the characteristics of a barrier-grid storage tube when used as a memory device for binary digital information.[2]

[1] A. S. Jensen, J. P. Smith, M. H. Mesner, and L. E. Flory, RCA Review, **9**, pp. 112–135, March, 1948.

[2] A similar tube was described by R. B. DeLano, Jr., Convention Record of the I.R.E., Part 3, pp. 125–130, 1954.

There is a wide latitude in the manner in which the tube may be operated. On the other hand, the speed of operation, the number of usable storage sites and the likelihood of error are interdependent. As the limiting storage capacity is approached, the likelihood of error increases sharply. The sources of variations in the output signals are described in this paper and a basis is given for determining the probability of error due to amplifier noise. Our discussion is concerned mainly with results obtained from experimental models of storage tubes. They utilize electrostatic beam focusing and deflection in order to permit high speed random access to any of the storage sites on the target.

A new type of reading and writing circuit was used in these investigations which is particularly suited to binary storage. The circuit allows signal detection at the storage target while maintaining good isolation between the input writing signal and the weak output reading signal.

II. OPERATING PRINCIPLES OF BARRIER-GRID TUBES

1. *Experimental Tubes*

A schematic drawing of the tube is shown in Fig. 1 and a photograph of the tube in Fig. 2. The electron gun, focusing system, and deflection

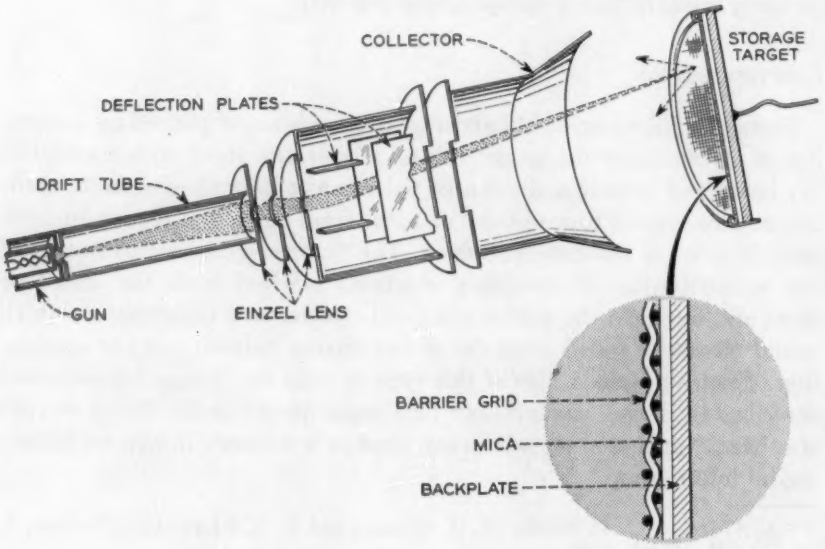


Fig. 1 — Schematic cross-section of a barrier-grid tube.

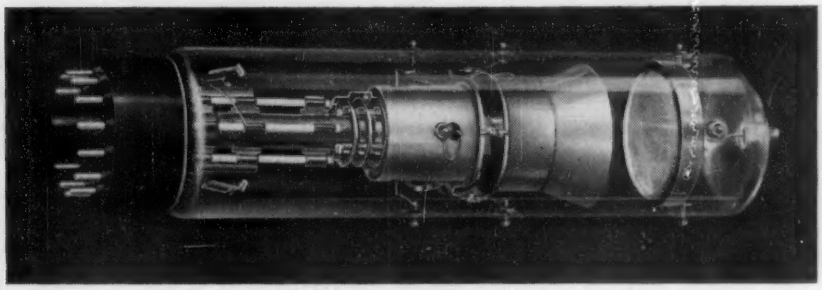


Fig. 2 — Experimental barrier-grid storage tube.

system are quite conventional and follow standard cathode-ray tube practice. The gun is a three-element immersion-lens structure of the type which produces a small crossover immediately in front of the cathode. The beam spreads outward from this crossover in an expanding cone and impinges on a limiting aperture at the end of a field-free drift tube. An Einzel type electron lens converges the trajectories of those central electrons which pass through the limiting aperture. This forms an image at the target of the first crossover in the gun. The electron beam at the target is roughly Gaussian in shape with a diameter such that the standard deviation, σ in Equation 1, of the current density distribution is about 0.002 inch. Ninety-five percent of the current will pass through a slit aperture 0.007 inch in width. The normal beam current is two microamperes and the accelerating potential is 1,000 volts.

An electron collector is placed just beyond the deflection plates to receive secondary electrons released at the storage target. The beam itself passes through a hole in the center of this collector. In one mode of operation, the signal is detected at this electrode, but in the target-reading mode, to be described later, the collector as a separate electrode is not required.

The storage target is a sandwich of three elements. A sheet of mica 0.001 inch thick is held between the barrier-grid and backplate. The grid is a woven mesh (wire cloth) of stainless steel with 400 wires per inch in each direction, the wires being 0.001 inch in diameter. These three elements are assembled to be in contact over the entire surface. The mica sheet, being a good insulator, will hold an electrostatic charge deposited on its surface for extended periods of time, thereby performing the storage function of the tube. The backplate is insulated from the grid and its potential may be varied to control the charge pattern laid down by the electron beam.

2. *Deposition of the Charge Pattern (Writing)*

In binary digital storage, we deposit either zero charge or a finite amount of charge at a large number of storage sites on the target. The useful area of a storage zone at each site is determined by the size of the beam. This is large enough to include, on the average, six small grid orifices. There is no attempt to register the beam with the fine structure of the barrier-grid and the beam may be allowed to fall on any portion of the target. In digital use, the deflection voltages must be rather precisely quantized so that the beam will strike repeatedly only at regularly spaced storage sites. The various sites may be designated or numbered in any desired way. Such a number or designation of a particular site is an "address" at which a digit may be stored.[3]

The mechanism of charge accumulation requires that the mica surface have a secondary emission ratio greater than unity, so that on the average each electron from the beam will release more than one secondary from the surface. If these secondaries are allowed to escape there will be a net loss of electrons from the surface which results in the accumulation of positive charge. If, on the other hand, these secondaries are returned to the surface, negative charge accumulates. When the potential of the mica surface is more negative than the grid, the secondaries escape and the potential of the mica surface rises. Conversely, when the mica has a potential substantially positive with respect to the grid, the second-

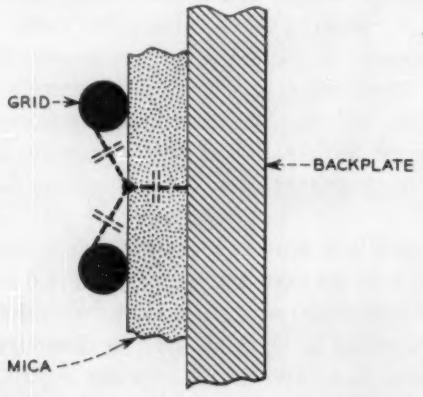


Fig. 3 — Capacitance distribution between the barrier grid, the dielectric surface, and the backplate.

[3] In most of our experiments the zones were spaced in a square lattice to obtain 16,384 sites in an area $1\frac{3}{8}$ inches on each side. This number is 2^{14} which allows us to designate the sites with a 14 digit binary number. Seven of these digits may be used to specify the vertical deflection voltage and seven the horizontal.

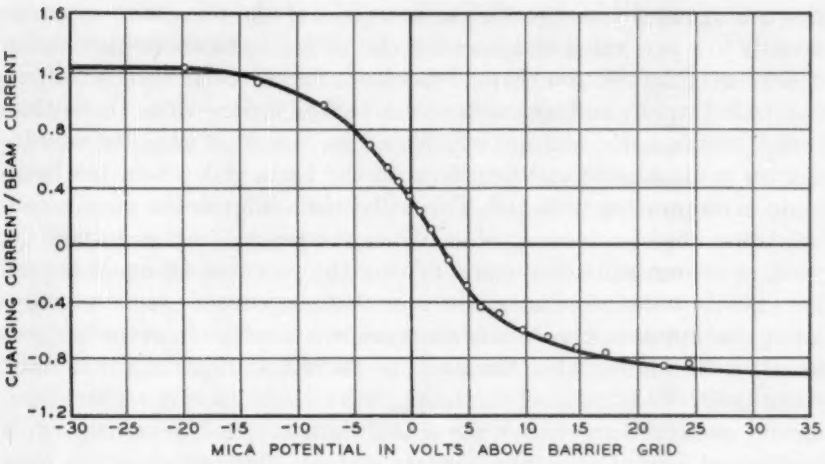


Fig. 4 — Charging curve at the mica surface showing the ratio of the charging current to the beam current as a function of the estimated potential difference between the dielectric surface and the barrier grid.

aries are returned and the potential drops. In each case the potential approaches an equilibrium value slightly more positive than the grid. At this potential the number of secondaries which escape is just equal to the number of primaries which arrive.

A degree of control over the potential of the mica surface may be obtained by varying the potential of the backplate behind the mica. One may visualize an equivalent circuit at the mica as indicated in Fig. 3. Here we picture capacitance distributed between the barrier-grid, the dielectric surface, and the backplate. Measurements indicate that on the average, 60 per cent of the backplate voltage change appears at the face of the mica with the barrier-grid at ground potential.

Fig. 4 shows the charging current as a function of the mica potential. This curve has a rounded characteristic partly because of the velocity distribution of the emitted secondary electrons. It is even more broadly rounded because of such effects as charging and discharging under the grid wires, redistribution of secondary electrons from the grid, and geometrical factors producing non-uniform electric fields. The curve of Fig. 4 was obtained by an indirect method using a broad, defocused electron beam sweeping over a large area.

In storing binary information, the backplate potential is switched between two voltages, usually 50 volts apart. Suppose that the beam has previously bombarded a spot until the equilibrium potential has been reached at the mica surface, and that we then apply a positive step-func-

tion voltage at the backplate. The front face of the mica rises instantaneously to a new value and negative charge begins to accumulate under the action of the electron beam. This causes the potential at the mica surface to fall rapidly and approach the same equilibrium value that it had before step-function was applied. We might now wait until the equilibrium is re-established and then turn off the beam and return the backplate to its previous potential. This will cause a drop in the mica potential below the equilibrium value. When the beam is again turned on, positive charge will accumulate, driving the potential up again toward the equilibrium value. During the time that negative charge is accumulating, the number of secondary electrons which escape from the target is less than the number when the mica has reached equilibrium. Conversely, when positive charge is accumulating, there is a temporary excess in secondary electron current over the equilibrium value. The output signal is obtained by detecting the secondary current. This output differs from the equilibrium value only during intervals of positive or negative charging of the mica. Such charging effects occur under the action of the beam during a short interval following any change in the backplate potential.

In describing the charging phenomena above, we have assumed a stationary beam. In actual service, we may *write* information at many sites by bombarding them successively for fixed time intervals. At each storage site we choose which of the two allowed backplate potentials to apply. At a later time, when we wish to *read* the information, we bombard the same spots but hold the backplate potential fixed at one of the two values. The signal types which are generated will be described in the next subsection.

3. *Reading Methods*

The reading signal is obtained by detecting the secondary emission current from the target. This may be accomplished either by measuring the net target current or by measuring the secondary current alone at the collector. These methods are called *target reading* and *collector reading*, respectively. For digital applications the target reading method offers several advantages over collector reading but collector reading is more convenient for image storage applications, which are not considered in this paper.

For digital operation, the beam is biased to cut-off except while performing either write or read operations at specific storage sites. Output pulses are generated by the tube whenever the beam is turned on. Either of two distinct types of pulses may be obtained while reading and additional types of output signal pulses may be generated during writing.

It is the function of the reading circuit and its associated amplifier and pulse discriminator to identify the stored binary digits.

During writing we bombard a number of spots, determining at each spot which of the two backplate potentials to apply. During reading, we return the beam to these spots but leave the backplate potential fixed at a potential which might be either of the two values used for writing. For reasons to be explained later, we choose to read at the more negative of the two values. Quite arbitrarily we designate an operation with the backplate positive as writing the binary digit *one*. An operation with the backplate at its more negative value is either a *write zero* operation or a *reading* operation. There are four possible cases, considering the backplate potential both at the time of the operation and at time of the last previous operation. These are: (a) positive at present, positive previously (b) negative at present, positive previously; (c) negative at present, negative previously; and (d) positive at present, negative previously. These operations are designated by the following names:

(a) *write one* on *one*
(b) *read one* (or *write zero* on *one*)
(c) *read zero* (or *write zero* on *zero*)
(d) *write one* on *zero*

In usual digital applications, an output signal is needed only during the (b) and (c) operations. The data presented in later sections deal only with these two types but in this section we will describe qualitatively all four signal types, which some modes of operation may require.

From previous discussions, it is apparent that operation (a) should give only the equilibrium secondary current leaving the target, operation (b) should give an excess secondary current, (c) should produce the equilibrium secondary current and (d) should produce less than the equilibrium secondary current.

The character of the output signal pulses differ in the collector and target reading methods. In collector reading, the secondary electrons arrive at the collector and generate a negative voltage proportional to the current. In target reading we detect the net current to the target which is the difference between the beam current and the secondary current.

Fig. 5 shows the waveforms associated with the collector reading method for each of the four possible cases. The top curve indicates that the beam is turned on for a short time to perform each operation, and is off between operations. In the same time sequence, we show the backplate potential switched to a positive value for *write one* and held at zero for reading. The third curve is an idealization of the collector current pro-

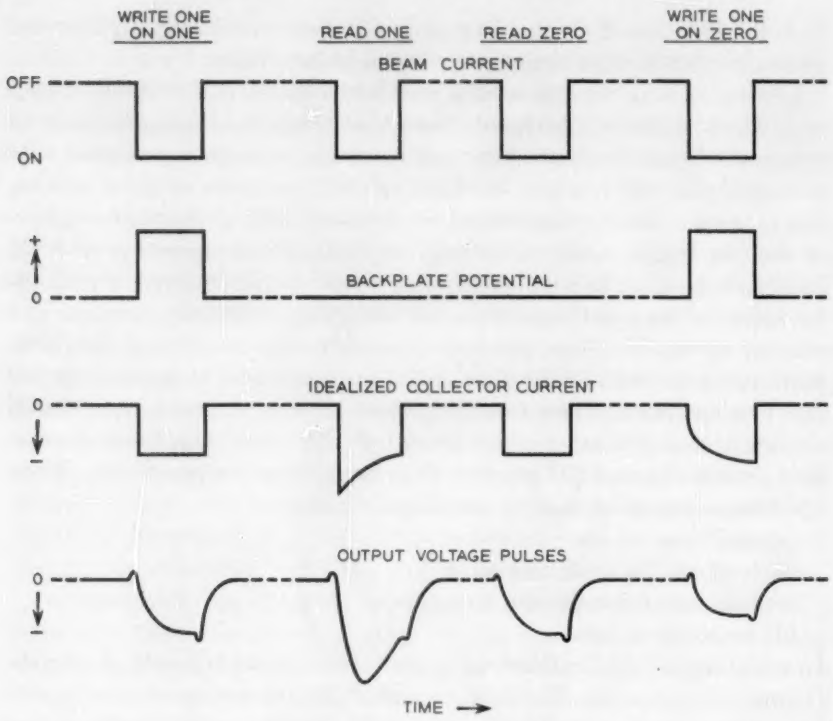


Fig. 5 — Signals obtained for four cases in binary signal storage, using collector reading.

duced in each operation. In cases (a) and (c) there is no change in the voltage condition of the surface and the current is constant at the equilibrium value. In the *read one* case, (b), the surface charges in the positive direction so that there is an excess current of secondary electrons to the collector. Because of rounded nature of the charging curve of Fig. 4, the excess current decreases as the surface charges toward equilibrium, producing the current pulse shown in Fig. 5. The pulse waveform depends on the amount of stored charge, the capacitance of the surface, and on the exact shape of the charging curve of the tube being investigated. In the case of *write one* on *zero*, (d), the surface charges in the negative direction so that less than the equilibrium current reaches the collector. The current gradually increases to the equilibrium value as the dielectric surface approaches the new equilibrium potential.

Experimental output voltage pulses found in our investigations are shown in the bottom curve of Fig. 5. The signal voltage is produced across a load resistor between the collector and ground. The rise and fall

times are determined primarily by the product of the collector capacitance and the load resistance. A *read one* output signal has the same polarity as a *read zero* signal, so that differentiation between the two types of output signals requires amplitude discrimination.

In collector reading, we detect only the electron current reaching the collector from the target structure. Some of the current from the target goes to the shield and is not utilized. Since the division of current between the shield and the collector depends on the position of the spot bombarded by the beam, a method which detects the total current leaving the target is desirable. This current can be detected by measuring the sum of the currents leaving all of the electrodes in the target structure. During the *write one* operation, however, a pulse of the order of 50 volts is applied to the backplate. Since the output signal is of the order of one millivolt, the reading circuit must be isolated as completely as possible from the writing circuit in order to prevent overloading the sensitive reading amplifier.

A target reading circuit which accomplishes this isolation is illustrated in Fig. 6. The basic circuit is a coaxial line which is coiled to form an inductance. The outer conductor is grounded at one end and is enlarged at the other end to surround the end of the tube and connect at

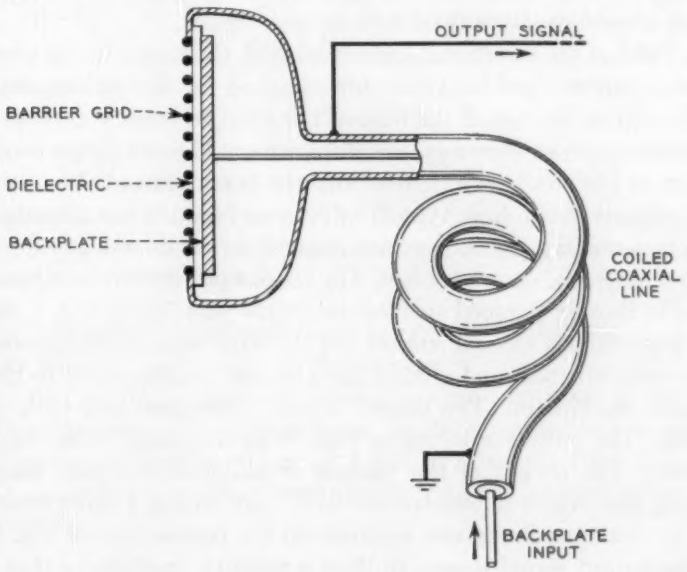


Fig. 6 — Schematic sketch of the circuit used for target reading in the barrier-grid storage tube. In these experiments, the coaxial line was wound on a ferrite core to obtain an adequately high inductance.

several places to the periphery of the barrier-grid. At the tube end, the outer conductor, which now includes the barrier-grid, completely encloses the backplate and shields it from ground. Upon application of a backplate pulse, current flows along the inner conductor to charge the barrier-grid to backplate capacitance, and ideally, all of this current should return through the outer conductor. Under these conditions no signal voltage should appear between the barrier-grid and ground because these currents should cancel each other and cause no voltage drop across the inductance. A net electronic current to the target, however, may divide between the inner and outer conductors in a parallel sense and thereby generate a signal whenever a change in this net current occurs. Ordinarily, the circuit is heavily damped with a shunt resistance from barrier-grid to ground to prevent ringing effects following signal generation.

Under equilibrium conditions, the effective secondary emission ratio of the target (including the grid and mica) is slightly less than one. Consequently, the *read zero* signal in this case is a small net negative current to the target within the tube. During intervals of charge accumulation on the mica, the net current will differ from this value. While accumulating positive charge (*read one*), the net current is somewhat positive and while accumulating negative charge (*write one* on *zero*) the current is more negative than the *read zero* signal.

Fig. 7 shows the waveforms associated with the target reading method. The beam current and backplate are switched for the various sequences as indicated at the top of the figure. The total secondary current which leaves the target is shown in the third curve. The net target current is the sum of the secondary current and the beam current. We also show output signals drawn from typical waveforms found in our investigations. The target reading circuit does not respond to the dc and low frequency components in the current pulses. The output pulses have an appearance similar to heavily damped sine waves. In the sketches in Fig. 7, the resonant frequency of the inductance and the stray capacitance between the barrier-grid structure and ground has a period roughly equal to the time the beam remains on. The output signals were measured with critical damping. The pulse shapes agree with those calculated from the circuit constants. The output of the reading amplifier should pass through a sampling gate which allows transmission only during a short time interval. The final gated response is shown on the bottom line of Fig. 7. The *read one* output signal is seen to have a polarity opposite to that of the *read zero* signal.

There are several advantages to the target reading scheme for binary digital storage. First, since the *zero* and *one* responses are of opposite

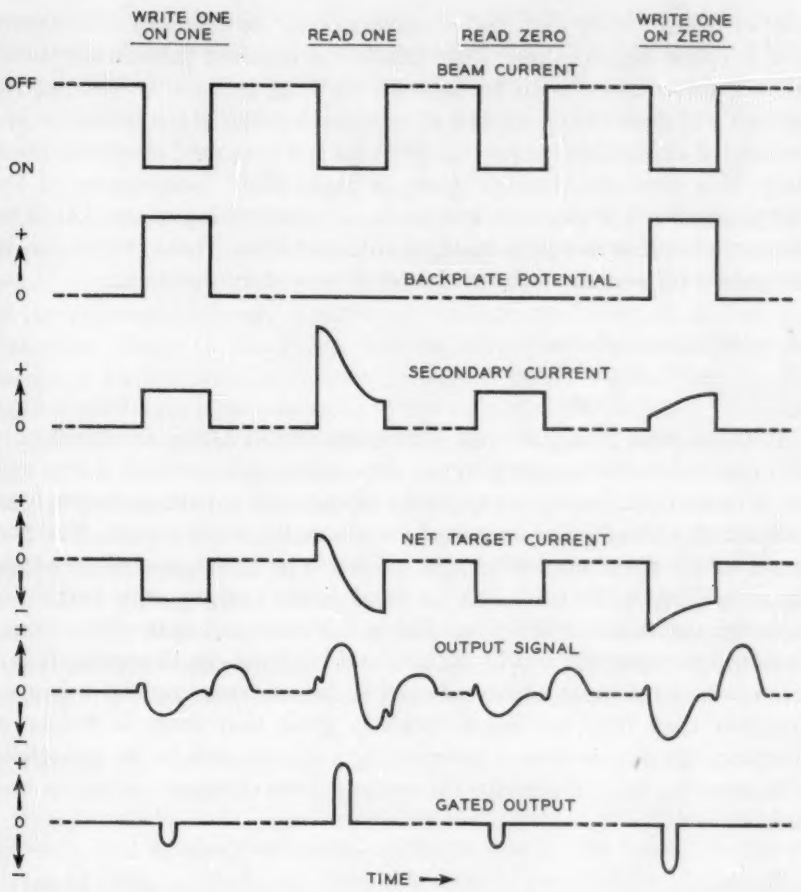


Fig. 7 — Currents and signals generated with the target-reading circuit of Fig. 6 for the four cases of Fig. 5.

polarity, the threshold decision level is near zero. This makes the output less sensitive to changes in beam current, which otherwise would require changes in the threshold level. Another advantage is that all of the secondary electron current from the target assembly is used for signal generation and there is no problem, as in collector reading, of obtaining uniform collection of secondaries from the various areas of the target surface.

4. *Regeneration*

In some conditions of operation, it is desirable to retain information in the tube after it is read. It is possible, by the use of external circuitry,

to rewrite any binary digit before moving on to the next spot. This operation is called regeneration. Regeneration is required periodically at all spots if information is to be retained for long periods. Otherwise, the pattern will deteriorate because of resistive leakage in the dielectric and because of neutralization by positive ions and scattered electrons. Similarly, if a random-access program is used, slight overlapping of the storage elements may result in erasure or false writing at one site if repeated operations are performed at adjacent sites. These effects can be minimized by periodic regeneration of all stored information.

III. SIGNAL VARIATIONS

1. *General*

With an ideal binary storage device one would expect to obtain only two types of output reading signal, depending upon whether a *zero* or a *one* is being read. These two types should be clearly distinguishable from each other, either by the magnitude or character of the signals. Furthermore, within their separate groups, all *one* type signals and all *zero* type signals should be identical. In a practical device working at its maximum capacity, the signals will not be ideal in this sense and appreciable variations will be observed within the *zero* and *one* type signal groups. It is a characteristic of binary systems that such variations can be tolerated provided that they are not sufficiently great that there is danger of interpreting a *zero* as a *one* or interpreting a *one* as a *zero*. In the remainder of this section we will describe the various types of signal variation and their causes.

2. *Shading*

Shading is a gradual variation in output signal over the surface of the storage target. Usually the signals are strongest at the center and weakest at the edge. One cause of shading is the finite angle of primary beam incidence for storage sites near the edges of the target. Appreciable charge accumulates under the grid wires where the escape of secondary electrons is somewhat inhibited. This effect appears both in target and collector reading schemes. When collector reading is used, there is additional shading because the fractional collection of secondary electrons is not uniform over the target surface, causing variations in the *zero* response as well as the *one* response.

A third cause of shading arises from defocusing of the electron beam by electrostatic deflection. This is a serious problem where wide angles of deflection are used. It is an inherent defect in electrostatic deflection and

cannot be eliminated by any degree of careful tube design. Pierce[4] shows that a set of deflection plates acts as a cylindrical converging lens whose focal length f is given approximately by the formula

$$\frac{1}{f} = \frac{2\theta^2}{l},$$

where θ is the angle of deflection and l is the length of the plates in the direction of the beam axis. Since the angle of deflection is directly proportional to the deflection voltage the focusing effect varies with the square of the deflection voltage. A correction voltage can be fed back from the deflection plates to the Einzel lens to compensate for this additional focusing. Experiments have been performed which verified that the required correction is proportional to the square of the deflection voltage. In the tubes used for these experiments, the maximum storage capacity can in fact be obtained only by the use of dynamic correction methods which compensate for this change in the focusing characteristics.

The shading variation across the target surface for both collector and target reading methods are shown in Fig. 8. Here we show the *zero* and *one* outputs along a line through the target center. The line is at a 45° angle to the deflection axes. In obtaining the data for these curves, a deflection defocusing correction was made. There is less shading with target reading than there is with the collector reading method.

3. *Texture*

Small variations in signal occur because the barrier grid is not sufficiently fine in mesh compared with the size of the beam. A slightly different signal is produced when the beam is centered over a grid hole than when it is centered at an intersection of grid wires. The beam thus resolves the mesh to some extent. Spotty variations also occur in the secondary emission characteristic of the barrier grid. The maximum variation in output pulse height from both of these effects is ±10 per cent.

4. *Blemishes*

Occasional variations in the storage and secondary emission characteristics are noted at spots on the mica surface. These may be caused by imperfections in the natural mica or by contamination with foreign matter. With care in selection of the mica and with additional care in tube

[4] J. R. Pierce, Theory and Design of Electron Beams, D. Van Nostrand and Co., 2nd Ed. 1954, pp. 41–46.

cleaning and assembly, blemishes are rarely observed which are of sufficient importance to cause errors in memory.

5. *Incomplete Erasure*

Whenever the potential at a spot is changed, the charging proceeds rapidly at first and then approaches the equilibrium condition in a manner similar to an exponential decay. In practice, one cannot wait an indefinite time to reestablish the equilibrium at each operation. This is further complicated by the non-uniform current density in the beam which does not allow the edges of the zones to become charged to equilibrium in normal operation. If the time of operation is prolonged, charge can continue to accumulate at the edges long after the center becomes saturated.

We shall now analyze an over-simplified model in order to obtain a

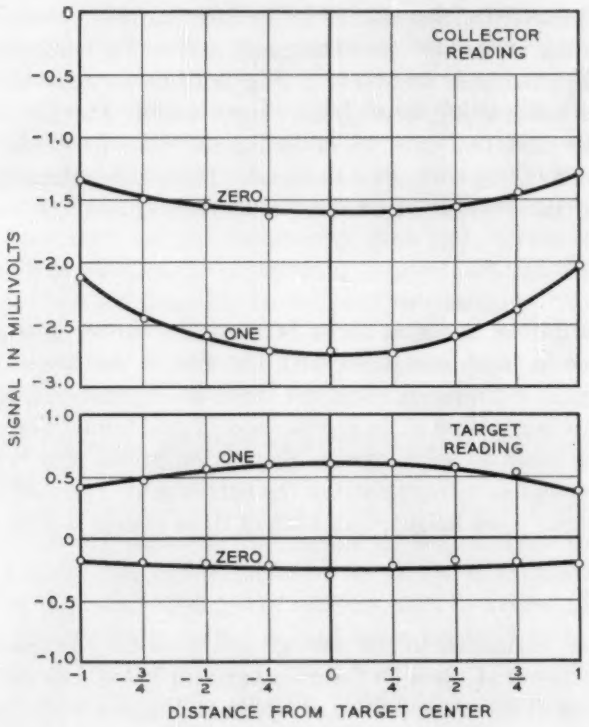


Fig. 8 — Shading characteristics for one microsecond reading and writing times. Data were taken along a diagonal through the center of the useful square area.

qualitative understanding of this behavior. We assume that the charging rate is proportional to the current density alone and that its magnitude is independent of the voltage difference between the grid and mica. The charging rate is positive for mica voltages below equilibrium and negative for voltages above. This is equivalent to assuming that the charging curve of Fig. 4 is a step-function, positive to the left of zero and negative to the right. We further assume that the beam has an ideal Gaussian variation of current density with radius (r) given by

$$J(r) = \frac{I_b}{2\pi\sigma^2} e^{-(r^2/2\sigma^2)} \tag{1}$$

where $J(r)$ is the current density, I_b is the beam current and σ is the measure of beam width.

As an initial condition, we assume that the surface has been brought to equilibrium with the backplate in its more negative state. We then raise the backplate to its positive state and perform a *write-one* operation at a fixed spot for an indefinite time. We wish to determine the character of the surface charge distribution which accumulates as a function of time. We also wish to know how this charge is removed in a subsequent reading operation. At each infinitesimal portion of the area, the charge accumulates in proportion to the local current density up to the saturation value. At first we obtain a simple Gaussian distribution growing linearly with time. After an interval which we designate as t_0, the center of the spot becomes saturated and the zone of saturation grows over an ever increasing area.

Fig. 9 illustrates the effect we are describing in this section. This figure shows several curves of charge density vs radius for various writing and

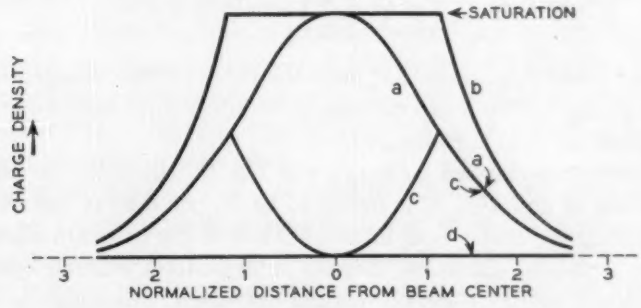


Fig. 9 — Charge density versus radius at a spot at various stages when the writing time and reading time are each 2 t_0. At curve (a) we have written for t_0 and at curve (b) we have finished writing at 2 t_0. At curve (c) we have read for an interval t_0 and at (d) we have finished reading after 2 t_0.

reading times. For these curves we assume that the charging current at the mica on reading is equal in magnitude but opposite in sign to that for writing.

When we write for a single time interval t_0, the charge distribution will be as shown by curve a. The center portion is just saturated. A read period of duration t_0 will completely discharge all points (curve d). When we write for an interval $2t_0$, the idealized charge distribution will be given by curve b where a large region is saturated near the center of the beam spot. If we now read for a period t_0, we will completely discharge a point at the beam center and leave an annular charged area as indicated by curve c. A second reading period of time t_0 will be required to completely discharge the surface. The discharge current during this second read period may be large enough to be interpreted as a *one*, producing an error signal.

We can analyze the situation to determine the seriousness of this effect. The radius, (r_1) of the saturated circular area after writing for an arbitrary time t, greater than t_0, can be found by noting that

$$J_{max}t_0 = J(r_1)t \tag{2}$$

where $J_{max} = I_b/2\pi\sigma^2$, the current density at the center of the Gaussian beam. Substituting for $J(r_1)$ from Eq. (1) and solving for r_1, we find

$$r_1 = \sigma\sqrt{2\,ln\,(t/t_0)} \tag{3}$$

The total charge deposited is

$$Q = J_{max}\pi r_1^2 t_0 + t\int_{r_1}^{\infty} 2\pi r J(r)\,dr \tag{4}$$

$$= I_b t_0\left(1 + ln\frac{t}{t_0}\right) \qquad (t > t_0) \tag{5}$$

Equations 3 and 5 indicate that both the radius of the charged area and the total charge continues to increase indefinitely in a logarithmic manner as the charging time is prolonged.

The charge represented by curve c of Fig. 9, which can be calculated with the aid of equation 5, is found to be 70 per cent of the charge removed during the first t_0 read interval. Thus, if the program allows more than one *write-one* operation without intermediate reading operations, an error could result. A second reading operation, which should indicate a *zero*, would give a signal which could be interpreted as a *one*.

The significance of the preceding discussion is that if we want high speed operation, we must allow a reading time of sufficient duration to remove substantially all of the charge deposited in previous *write-one*

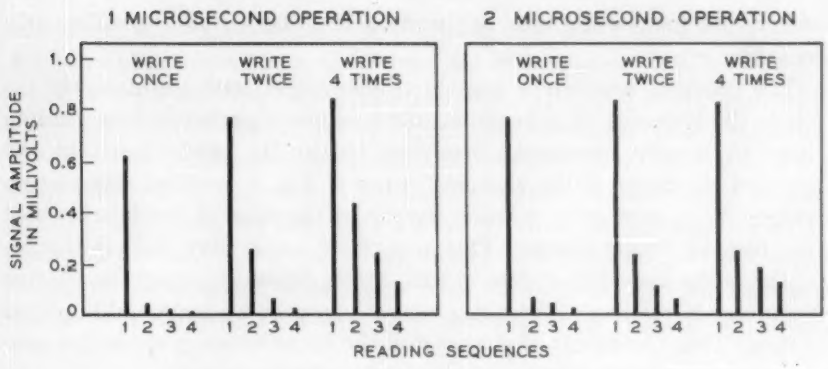


Fig. 10 — Successive reading signals obtained after repeated write-one operations. The value plotted is the difference between the read output and the ideal read zero signal.

operations. This suggests that a reading operation should precede each *write-one* operation.

The incomplete erasure effect is less significant if reading and writing times are both made much longer than t_0. Fig. 10 shows the result of increasing the normal operating times. Here we show actual output signals obtained from a tube for the first and subsequent reading operations. The signal amplitudes are shown for several successive *read* operations following one or more *write-one* periods. Two sets of data are shown, one for one microsecond operating times and the other for two microsecond times. The spurious output signals found during the second reading period are considerably reduced if longer operating times are used. For example, the second read signal after four *write-one* operates is about 60 per cent of the normal read signal when an operating time of one microsecond is used. When the time is two microseconds it is only 30 per cent of the normal read signal. However, the improvement obtained by using two microsecond pulses is expensive in that the speeds are reduced and interference effects from overlapping distributions are increased, requiring a larger spacing between sites.

Another method which might be used to minimize this effect would be to use longer times for reading than we do for writing. The same purpose can be accomplished with equal reading and writing times if the reading operation is more effective in discharging the storage surface than is the writing action in depositing charge. The charging curve of Fig. 4 shows a greater charging rate in the positive than in the negative direction. This is one of the reasons why we choose to read at the more negative backplate voltage condition. Otherwise, we would be required to use a

longer time for reading than for writing in order to obtain sufficiently complete erasure.

The previous analysis is greatly oversimplified but qualitatively explains the necessity of a program which allows sufficiently long reading times to remove previously deposited charge. In actual practice, the rounded character of the charging curve of Fig. 4 modifies this picture considerably, especially in those regions at the edge of the beam where the charging action is weak. This may cause a very slow drift in the potential of the dielectric surface in this region, depending upon the relative frequency of reading and writing operations over an appreciable period of time. Thus the extent of charging in the areas between spots in a spacial array will depend upon the particular storage program.

The output signals on reading are affected to some extent by the presence of charge in the intermediate regions, so that we can expect a certain amount of variability in the output depending upon the recent history of *write zero* and *write one* operations. One observed consequence is that we obtain larger signals from stored *ones* as the frequency of *write one* operations increases in comparison with *write zero* operations.

If we alternately *write one* and *read* at a particular site over a long period of time, the read output signal may be as much as 40 per cent larger than that obtained when the frequency of *write one* operations is very low. Thus, we find a significant source of signal variation caused by incomplete erasure. These effects may be minimized by restricting the program so that a reading operation precedes each *write one* operation or by allowing an adequately long reading time compared to the writing time.

6. *Overlap*

Another consequence of the Gaussian distribution of current density is that there will be a slight overlap of the storage elements, as defined by the beam. If two spots overlap slightly, repeated reading at one spot will tend to erase slowly a *one* signal stored at the other. Similarly, repeated writing of *ones* at a given spot will gradually cause a *one* to appear at an adjacent spot which is initially at *zero*. A measure of the seriousness of this effect is the "read-around-ratio," which is defined as the number of times that reading, writing, erasing, or regenerating operations may be performed at one spot without destroying information at an adjacent spot. Read-around-ratio and resolution (i.e., the number of storage sites per unit area) are interdependent, so that less interaction between spots can be obtained by reducing the number of storage sites. Conversely, an increase in the number of storage sites reduces the read-around-ratio. The relationship between read-around-ratio and distance between storage

sites was found experimentally in a typical tube and is shown in Fig. 11. An amplitude degradation of 25 per cent of maximum was used in the measurements for the read-around criterion. The solid curve represents the relationship which would be expected on a simple theory based on the Gaussian distribution of current density. This curve is drawn for a σ of 0.0026″ which agrees well with independent measurements of spot size. The difference between the theoretical values and the data for larger distances between storage sites may be due to redistribution of secondary electrons, to scattering of electrons from the edges of the gun apertures and to electron optical effects such as lens aberrations and space charge.

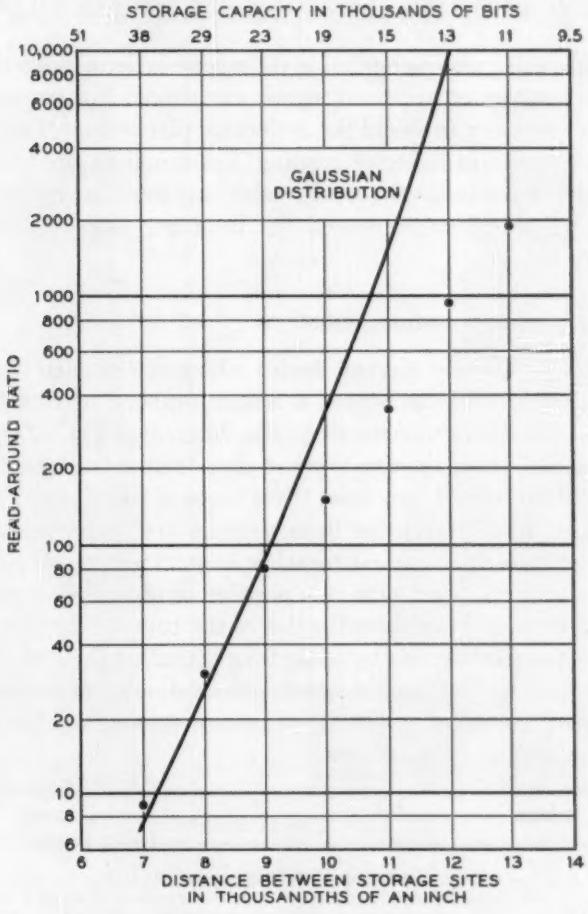


Fig. 11 — Read-around ratio and storage capacity as a function of the distance between adjacent storage sites.

7. *Noise*

Any perturbing noise phenomenon will affect the uniformity of the output signal. The beam current contains the usual fluctuation noise arising from the random times of emission of electrons at the cathode and from the velocity distribution of emission. To this must be added some additional noise because secondary electrons are emitted in groups. There are, however, about 10^7 electrons in each signal pulse and the probability of this number varying significantly because of shot effect is entirely negligible. The random noise generated in the reading amplifier can, however, be a significant source of error if the amplifier is poorly designed or if the signal strength is too weak. The effects of amplifier noise will be discussed later.

Man-made noise and interference pickup of various kinds can also be a significant source of apparent signal variations. For example, it has been found necessary to shield the deflection plates from the collector to reduce interference in collector reading. Inasmuch as the output signal is of the order of one millivolt, considerable care must be used to eliminate all sources of interference caused by improper shielding, circulating ground currents, etc.

IV. OUTPUT SIGNAL CONSIDERATIONS

In any large capacity storage device where the storage sites are distributed in space we must expect a certain amount of variation in the output signals from the various sites. The diagram of Fig. 12 might represent the magnitudes of the two types of signals from each site in a binary store. Ideally, we would like these two groups of signal amplitudes to be well separated from each other in magnitude and to have the members of each group closely bunched together about their respective average values. In the barrier-grid tube it is possible to obtain such performance if one is overly conservative in the use of the tube by keeping the spots large and well separated and by using long operating periods. If the spots are more closely spaced, higher speeds are used, and random-access programs applied, the variations in signal become greater and the separation between the groups becomes less.

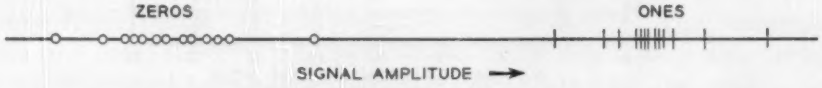


Fig. 12 — A possible set of signals from each site in a binary store, showing a range of signal variations.

It has been found convenient to express the signal characteristics and uniformity by means of a signal separation chart as shown in Fig. 13. On this chart, the output signal voltage is the abscissa and the ordinate represents numbers of storage sites. A square-root scale is used for the ordinate to emphasize the region of minimum signal separation. We plot two curves of the "integral" type. The curve on the left represents the total number of spots which give a *zero* response voltage greater than the abscissa; the curve on the right is the number giving a one response less than the abscissa. Each of these curves is measured using the program of operation most likely to produce an error. For example, the program of operation most likely to produce an error for a written *one* is as follows: Store *ones* at alternate spots in the array, read the remaining spots some

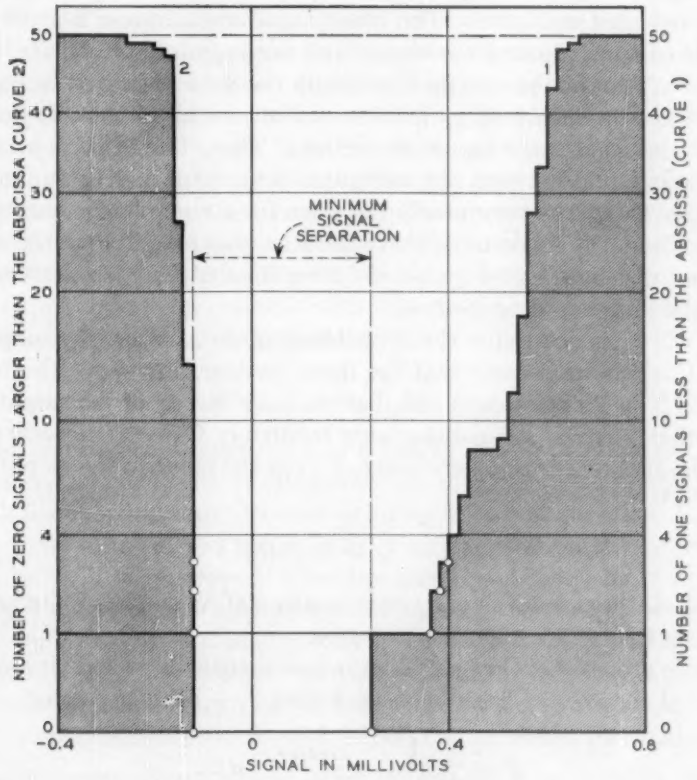


Fig. 13 — Signal separation chart for 50 scattered spots on the barrier grid storage tube. Curve 1 shows the "one" signals obtained after 100 reading operations are performed adjacent to the written spots. Curve 2 shows the "zero" signals obtained after 100 write one read sequences are applied to adjacent spots.

large number of times, then return and read the original group. We then record the amplitudes of the final reading signals and plot the results on the right side of the signal separation chart. The corresponding pessimistic program for the *zero* group is to *write zero* on alternate spots, then write alternate *zeros* and *ones* on the remaining spots for a large number of cycles, then return and read the original group. Plotting the distribution of the *zero* type outputs obtained completes the picture on the signal separation chart. The minimum separation between these curves gives an indication of the likelihood of error in a random-access type of service.

The signal separation chart is a representation of the total effect of the signal variations discussed in Section III. The effects of amplifier noise, however, are greatly reduced by our method of performing the tests, since the signal obtained for each spot is the average of a large number of repeated operations. The other signal variations either are systematic (shading, incomplete erasure and overlap) or have definite limits (texture). The storage sites used to obtain the data presented in Fig. 13 were chosen in such a way that the worst conditions contributing to systematic signal variations were included. Thus, the signal separation which was found between the maximum *zero* signal and the minimum *one* signal should be very nearly the same for a much larger number of storage sites. The minimum signal separation was found to be 0.36 millivolts out of a total signal spread of 0.90 millivolts. The data were taken using the target-reading method.

We will next determine the probability of error caused by amplifier noise. For this we assume that the input resistor is R ohms, the bandwidth is B cycles per second and that the noise factor[5] of the amplifier is F. Further we must assume that only random or Gaussian noise is present. The apparent rms noise voltage, V_n, at the input to the amplifier is given by[5]

$$V_n = \sqrt{4FRkT_0B}\,, \tag{6}$$

where T_0 is the standard noise temperature, 290°K, and k is Boltzmann's constant, 1.38×10^{-23} joules per degree.

A basic property of Gaussian fluctuations is that the probability at any instant of the voltage lying between V and $V + dV$ is given by[6]

$$P = \frac{1}{V_n\sqrt{2\pi}}\, e^{-(V^2/2V_n^2)}\, dV\,. \tag{7}$$

---

[5] Stanford Goldman, Frequency Analysis, Modulation and Noise, McGraw-Hill, 1948.

[6] Ibid, p. 206.

Integrating from V to infinity gives the probability, $P_{>V}$, that the noise voltage will be greater than V. This gives

$$P_{>V} = \frac{1}{2}\left(1 - \frac{2}{\sqrt{\pi}} \int_0^{V/\sqrt{2}V_n} e^{-x^2}\, dx\right) \tag{8}$$

The second term inside the parenthesis is the familiar error integral given in Pierce's integral tables.[7] Suppose, now, that V represents the voltage difference between a given signal and discrimination level (threshold). The probability of error, due to noise, for *that* signal is given by equation (8), substituting for V from equation (6).

It is interesting to assume some reasonable numbers for the amplifier characteristics and determine the minimum allowed signal voltage V to obtain an arbitrarily chosen probability of error. We assume a program such that 10^6 operations are performed per second. Thus a probability of error of 4×10^{-13} corresponds to one mistake per month of continuous operation. We also choose an amplifier with a bandwidth of two megacycles per second and a noise factor of 2. A typical input resistor has a value of 2,000 ohms. Under these conditions we calculate from (8) that a minimum signal of 80 microvolts measured from threshold is allowed. A separation of 360 microvolts was found on the signal separation chart. There is, therefore, a negligible probability of error due to amplifier noise alone. Variations in the voltages applied to the storage tube or drift in the output circuit can combine with noise to produce errors unless care is taken to provide stable external circuitry.

V. SUMMARY

The barrier-grid storage tube as a binary digital memory device requires special conditions of operation. The margin of safety with regard to the likelihood of error depends upon the speed of operation, the type of program applied, and the spacing of storage sites on the target. The probability of error depends upon the minimum separation of *zero* and *one* type output signal amplitudes and upon the noise factor of the output amplifier. This probability may be made vanishingly small if the method of operation is conservatively chosen. Careful investigation of all these factors indicates that with a laboratory model of a storage tube, reliable operation is possible with 16,000 bits of stored information per

[7] B. O. Pierce, A Short Table of Integrals, Gin and Co., pp. 116–120, 1929, For large values of the argument we can use the semiconvergent series.

$$\frac{2}{\sqrt{\pi}} \int_0^y e^{-x^2}\, dx = 1 - \frac{e^{-y^2}}{y\sqrt{\pi}}\left(1 - \frac{1}{2y^2} + \frac{3}{(2y^2)^2} - \cdots\right)$$

tube, a read-around-ratio greater than 100, and a characteristic time of about one microsecond per operation.

IV. ACKNOWLEDGMENT

R. W. Sears was directly responsible for the earlier exploratory development phases of barrier-grid storage tubes in our laboratories and his continued guidance has been invaluable. H. C. Jonas carried out the mechanical phases of this work.

Distortion in Feedback Amplifiers

By R. W. KETCHLEDGE

(Manuscript received May 17, 1955)

Distortion in feedback amplifiers and other non-linear circuits is analyzed for the case where the magnitude and phase of the feedback varies with frequency. The analysis is limited to cases where the distortion products are small compared to the fundamentals and where the non-linear element can be described by a power series having only a few terms. However, many practical amplifiers are adequately described by the analysis. Formulae are derived for a number of third-order products and their dependence upon various feedbacks at second order frequencies is demonstrated.

INTRODUCTION

Distortion in feedback amplifiers has previously been studied for the case where the feedback is independent of frequency.[1] However, in many practical cases, the variation of feedback with frequency produces significant deviations from this simple theory.[2] The present analysis takes into account the magnitude and phase of the feedback at all frequencies in determining the amount of any particular modulation product. This analysis has proved useful in the design of amplifiers for the L3 coaxial carrier system[3, 4] as well as in the analysis of a number of non-linear circuits. The method is most useful in cases where the distortion products are small compared to the fundamental signals and where the non-linear element can be described by a power series having only a few terms. More complex cases can be treated but the labor involved is appreciably greater. However, many practical feedback amplifiers are adequately described by the analysis and, in addition, some understanding is obtained as to the mechanisms involved. In particular, the dependence of third order distortions on the feedbacks at second order frequencies is demonstrated and formulae are obtained.

THE PROBLEM

When a signal is sent through a non-linear element, such as a vacuum tube, the output can usually be described as a power series of the input

signal. This is especially true in wideband amplifiers where plate load impedances are low and plate current is determined largely by grid-cathode voltage. Often the non-linear element is contained within a feedback loop such that a portion of the output is returned to the input. In this situation the total input contains a power series of the original input and the situation is considerably complicated. It is well-known,[1, 2] for example, that third harmonics can be produced not only by the cube term of the power series but, with feedback, by the square term. The square term produces second harmonic output which, after being fed back, mixes again with the fundamental to form (via the square again) third harmonics. Thus, the third harmonic output becomes dependent to some degree on the feedback at the second harmonic. This relationship becomes somewhat more complex when several fundamental inputs are present simultaneously but, in general, the output of a particular third order product depends on the feedback at, at least, some of the second order product frequencies. Within the limitations of the simplifying assumptions used, the present analysis evaluates these relationships.

THE METHOD AND THE ASSUMPTIONS

It is assumed that the non-linear element can be described by a power series of the form

$$i = a_1 e + a_2 e^2 + a_3 e^3 + a_4 e^4 + \cdots . \tag{1}$$

In a circuit such as shown on Fig. 1, e represents incremental grid-cathode voltage and i, incremental plate current of a vacuum tube. If, as shown on Fig. 1, a fraction of the output is returned to the input, then the grid-

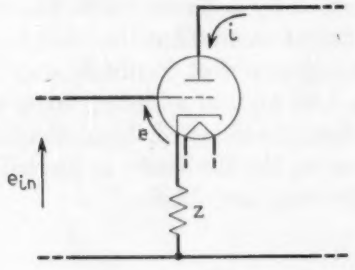


Fig. 1 — Feedback Amplifier Equivalent Circuit.

cathode voltage becomes the difference between the applied and fed back signals or,

$$e = e_{in} - iZ \tag{2}$$

While Z has the dimension of impedance, it is not limited to the ± 90 degrees of an ordinary two terminal impedance. In practice, $a_1 Z$ usually represents the loop gain, $(\mu\beta)$, of the feedback amplifier. Thus, Fig. 1 is used to represent the voltage and current relationships of any feedback amplifier.

Having expressed the output current as a power series of the grid-cathode voltage (1) and having expressed the grid cathode voltage as the sum of the input and fedback voltages, (2), the next step is to combine these expressions. The method consists of first expressing the input voltage, e_{in}, in terms of its cosinusoidal components. The grid-cathode voltage is expressed as a Fourier series of cosines having complex coefficients. The frequencies in the series represent all combinations and harmonics of the frequencies present in the input signal itself. The coefficients of the power series for the current, a_1, a_2, a_3, etc., are known from the tube characteristics. The problem concerns itself initially with finding the unknown coefficients for the Fourier series representation of the grid-cathode voltage.

The method is to first insert the Fourier series for e in the power series to find a Fourier series for i in terms of the coefficients for e. Then, equating the two sides of the equation, frequency by frequency, one obtains a set of simultaneous equations which can be solved for the coefficients of e and in turn for the coefficients of i. It so happens that with the assumptions used here the equations can be solved individually.

The chief difficulty of the method resides in the fact that both the power series and Fourier series are infinite and therefore a rigorous solution is impractical. Fortunately, simplifying assumptions permit limiting the series in many practical cases without seriously degrading the accuracy. The assumptions used here are as follows:

1. All distortion products are small compared to the fundamentals.
2. Third order distortion products are small compared to second order products.
3. The device non-linearity is adequately described by a simple three term power series. Fourth order and higher powers are neglected.
4. Frequency components representing fourth order and higher interactions are neglected.

"SINGLE-FREQUENCY INPUT"

In the single-frequency case the input signal, e_{in}, is given simply by

$$e_{in} = A \cos \alpha t \tag{3}$$

Since the input is a single frequency, the Fourier series representation of the grid-cathode voltage, e, is known to contain merely the harmonics of the input frequency. Thus,

$$e = \sum_{n=0}^{n=\infty} k_n \cos n\alpha t \tag{4}$$

or

$$e = k_0 + k_1 \cos \alpha t + k_2 \cos 2\alpha t + k_3 \cos 3\alpha t \cdots \tag{5}$$

The next step is to insert (5) in (1) limiting (1) to the first three terms as follows:

$$i = a_1 e + a_2 e^2 + a_3 e^3 \tag{6}$$

In doing this it must be remembered that k_n is complex and represents a complex voltage. Therefore, the products are not the ordinary result of the product of two complex numbers where the magnitudes multiply and and the angles add. Rather, these products, representing complex numbers pertaining to impedance, voltages or currents of different, or the same, frequencies are formed by using the conjugate of any coefficient whose frequency subtracts in the formation of the product frequency. For example,

$$k_a k_b \cos A \cos B = \tfrac{1}{2} k_a k_b \cos (A + B) + \tfrac{1}{2} k_a \bar{k}_b \cos (A - B)$$
$$(\bar{\ }) = \text{conjugate}, \qquad A > B \tag{7}$$

If this rule were not followed at least the phases of the products would be incorrect.

Performing the operation for e^2 and combining terms of the same frequency yields,

$$\begin{aligned} e^2 = k_0^2 + \frac{k_1\bar{k}_1 + k_2\bar{k}_2 + k_3\bar{k}_3}{2} + (+2k_0k_1 + \bar{k}_1k_2 + \bar{k}_2k_3) \cos \alpha t \\ + (2k_0k_2 + \tfrac{1}{2}k_1^2 + k_3\bar{k}_1) \cos 2\alpha t + (2k_0k_3 + k_1k_2) \cos 3\alpha t \\ + (k_1k_3 + \tfrac{1}{2}k_2^2) \cos 4\alpha t + k_2k_3 \cos 5\alpha t + \tfrac{1}{2}k_3^2 \cos 6\alpha t \end{aligned} \tag{8}$$

A similar operation for e^3 yields

$$\begin{aligned}
e^3 &= k_0^3 + \tfrac{1}{2}[k_0k_1\bar{k}_1 + k_0k_2\bar{k}_2 + k_0k_3\bar{k}_3] + \tfrac{1}{2}Re[2k_0(k_1^2 + k_2^2 + k_3^2) \\
&+ k_1k_2\left(\bar{k}_1 + \frac{k_1}{2}\right) + \tfrac{1}{2}k_3(k_1k_2 + k_1\bar{k}_2 + \bar{k}_1k_2)] \\
&+ ((+3k_0^2k_1 + 3k_0\bar{k}_1k_2 + 3k_0\bar{k}_2k_3 + \tfrac{3}{4}k_1^2\bar{k}_1 + \tfrac{3}{4}\bar{k}_1^2k_3 + \tfrac{3}{2}k_1k_2\bar{k}_2 \\
&+ \tfrac{3}{2}k_1k_3\bar{k}_3 + \tfrac{3}{4}k_2^2\bar{k}_3)) \cos \alpha t \\
&+ ((+3k_0^2k_2 + \tfrac{3}{2}k_0k_1^2 + 3k_0\bar{k}_1k_3 + \tfrac{3}{2}k_1\bar{k}_1k_2 + \tfrac{3}{2}k_1\bar{k}_2k_3 \\
&+ \tfrac{3}{4}k_2^2\bar{k}_2 + \tfrac{3}{2}k_2k_3\bar{k}_3)) \cos 2\alpha t \qquad (9) \\
&+ ((3k_0^2k_3 + 3k_0k_1k_2 + \tfrac{1}{4}k_1^3 + \tfrac{3}{2}k_1\bar{k}_1k_3 + \tfrac{3}{4}\bar{k}_1k_2^2 \\
&+ \tfrac{3}{2}k_2\bar{k}_2k_3 + \tfrac{3}{4}\bar{k}_3k_3^2)) \cos 3\alpha t \\
&+ ((\tfrac{3}{2}k_0k_2^2 + 3k_0k_1k_3 + \tfrac{3}{4}k_1^2k_2 + \tfrac{3}{2}\bar{k}_1k_2k_3 + \tfrac{3}{4}\bar{k}_2k_3^2)) \cos 4\alpha t \\
&+ ((3k_0k_2k_3 + \tfrac{3}{4}k_1^2k_3 + \tfrac{3}{4}k_1k_2^2 + \tfrac{3}{4}\bar{k}_1k_3^2)) \cos 5\alpha t \\
&+ ((\tfrac{3}{2}k_0k_3^2 + \tfrac{3}{2}k_1k_2k_3 + \tfrac{1}{4}k_2^3)) \cos 6\alpha t \\
&+ (\tfrac{3}{4}k_1k_3^2 + \tfrac{3}{4}k_2^2k_3) \cos 7\alpha t + \tfrac{3}{4}k_2k_3^2 \cos 8\alpha t + \tfrac{1}{4}k_3^3 \cos 9\alpha t
\end{aligned}$$

We now introduce the assumptions. Specifically,

$$k_1 \gg k_2 \gg k_3\,, \qquad k_1 \gg k_0 \gg k_3 \qquad (10)$$

and we neglect k_4, k_5, etc. The reduction in labor is apparent from inspection of equations 8 and 9. This simplifies equations (5), (8) and (9) as follows:

$$e = k_0 + k_1 \cos \alpha t + k_2 \cos 2\alpha t + k_3 \cos 3\alpha t \qquad (11)$$

$$e^2 = \tfrac{1}{2}k_1\bar{k}_1 + (2k_0k_1 + \bar{k}_1k_2) \cos \alpha t + \tfrac{1}{2}k_1^2 \cos 2\alpha t + k_1k_2 \cos 3\alpha t \qquad (12)$$

$$\begin{aligned}
e^3 = \tfrac{1}{2}k_0\bar{k}_1k_1 + Re\left[k_0k_1^2 + \tfrac{1}{2}k_1k_2\left(k_1 + \frac{\bar{k}_1}{2}\right)\right] + \tfrac{3}{4}k_1^2\bar{k}_1 \cos \alpha t \\
+ (\tfrac{3}{2}k_0k_1^2 + \tfrac{3}{2}k_1\bar{k}_1k_2) \cos 2\alpha t + \tfrac{1}{4}k_1^3 \cos 3\alpha t
\end{aligned} \qquad (13)$$

From (2) and (6) we know

$$e = A \cos \alpha t - Z[a_1e + a_2e^2 + a_3e^3] \qquad (14)$$

which may be written as

$$A \cos \alpha t = (1 + Za_1)e + Za_2e^2 + Za_3e^3 \qquad (15)$$

Using the values for e, e^2, e^3, given by (11), (12) and (13), in (15) and

solving, frequency by frequency, yields for dc

$$0 = (1 + Z_0a_1)k_0 + Z_0a_2\tfrac{1}{2}k_1\bar{k}_1 + Z_0a_3\left(\tfrac{1}{2}k_0\bar{k}_1k_1 + Re\left[k_0k_1^2 + \tfrac{1}{2}k_1k_2\left(k_1 + \frac{\bar{k}_1}{2}\right)\right]\right) \tag{16}$$

For α

$$A = (1 + Z_1a_1)k_1 + Z_1a_2(2k_0k_1 + \bar{k}_1k_2) + \tfrac{3}{4}Z_1a_3k_1^2\bar{k}_1 \tag{17}$$

For 2α

$$0 = (1 + Z_2a_1)k_2 + \tfrac{1}{2}Z_2a_2k_1^2 + Z_2a_3(\tfrac{3}{2}k_0k_1^2 + \tfrac{3}{2}k_1\bar{k}_1k_2) \tag{18}$$

For 3α

$$0 = (1 + Z_3a_1)k_3 + Z_3a_2k_1k_2 + \tfrac{1}{4}Z_3a_3k_1^3 \tag{19}$$

In order for assumptions 1 and 2 to be met, namely,

$$k_1 \gg k_0\,, \qquad k_2 \gg k_3\,, \qquad k_1 \gg k_2\,, \qquad k_0 \gg k_3\,, \tag{20}$$

it is necessary that

$$a_1e \gg a_2e^2 \gg a_3e^3 \tag{21}$$

or

$$a_1 \gg a_2e \gg a_3e^2. \tag{22}$$

Therefore, since

$$a_2k_1^2 \gg a_3k_1^2k_0 \qquad \text{or} \qquad a_3k_1^2k_2 \tag{23}$$

Equation 18 can be solved directly for k_2 yielding

$$k_2 = -\frac{1}{2}\frac{a_2Z_2}{1 + a_1Z_2}k_1^2 \tag{24}$$

and likewise

$$k_0 = -\frac{a_2Z_0}{1 + a_1Z_0}k_1\bar{k}_1 \tag{25}$$

This procedure avoids a simultaneous solution of (16) to (19). Since, for small distortion.

$$k_1 \approx \frac{A}{1 + a_1Z_1} \tag{26}$$

and having k_0 and k_2, one obtains

$$k_0 = -\frac{a_2}{2}\left(\frac{Z_0}{1 + a_1Z_0}\right)\left(\frac{A}{1 + a_1Z_1}\right)\left(\overline{\frac{A}{1 + a_1Z_1}}\right) \tag{27}$$

$$k_1 = \frac{A}{1 + a_1Z_1} - Z_1\left[\frac{3a_3}{4} - \frac{a_2^2Z_0}{1 + a_1Z_0} - \frac{a_2^2Z_2}{2(1 + a_1Z_2)}\right] \cdot \left(\frac{A}{1 + a_1Z_1}\right)^2\left(\overline{\frac{A}{1 + a_1Z_1}}\right) \tag{28}$$

$$k_2 = \frac{a_2}{2}\left(\frac{Z_2}{1 + a_1Z_2}\right)\left(\frac{A}{1 + a_1Z_1}\right)^2 \tag{29}$$

$$k_3 = \frac{Z_3}{1 + a_1Z_3}\left[\frac{a_3}{4} - \frac{a_2^2Z_2}{2(1 + a_1Z_2)}\right]\left(\frac{A}{1 + a_1Z_1}\right)^3 \tag{30}$$

The corresponding values for the output currents are readily obtained as,

$$i_0 = \frac{a_2}{2(1 + a_1Z_0)}\left(\frac{A}{1 + a_1Z_1}\right)\left(\overline{\frac{A}{1 + a_1Z_1}}\right) \tag{31}$$

$$i_1 = \left[\frac{a_1A}{1 + a_1Z_1} + \left(\frac{3a_3}{4} - \frac{a_2^2Z_0}{1 + a_1Z_0} - \frac{a_2^2Z_2}{2(1 + a_1Z_2)}\right) \cdot \left(\frac{A}{1 + a_1Z_1}\right)^2\left(\overline{\frac{A}{1 + a_1Z_1}}\right)\right] \cos \alpha t \tag{32}$$

$$i_2 = \frac{a_2}{2(1 + a_1Z_2)}\left(\frac{A}{1 + a_1Z_1}\right)^2 \cos 2\alpha t \tag{33}$$

$$i_3 = \frac{1}{1 + a_1Z_3}\left[\frac{a_3}{4} - \frac{a_2^2Z_2}{2(1 + a_1Z_2)}\right]\left(\frac{A}{1 + a_1Z_1}\right)^3 \cos 3\alpha t \tag{34}$$

The expression for the fundamental output current includes the third order distortion of fundamental frequency. This is often viewed as a gain change and, in order to keep the polarity positive, will be expressed as expansion or increase in gain. The expansion is defined here as the ratio of the gain to the gain at small signal levels. The gain at small signal levels is obviously found by

$$A \to 0, \qquad i_1 \to \frac{a_1A}{1 + a_1Z_1} \cos \alpha t \tag{35}$$

and therefore,

$$\text{Expansion} = \frac{i_1}{\dfrac{a_1A}{1 + a_1Z_1} \cos \alpha t} \tag{36}$$

Thus,

$$\text{Expansion} = 1 + \left[\frac{3a_3}{4a_1} - \frac{a_2^2 Z_0}{a_1(1 + a_1 Z_0)} - \frac{a_2^2 Z_2}{2a_1(1 + a_1 Z_2)}\right] \cdot \left(\frac{A}{1 + a_1 Z_1}\right)\left(\frac{A}{1 + a_1 Z_1}\right) \tag{37}$$

It should be pointed out that both this solution and the others to follow can be applied to other non-linear circuits besides feedback amplifiers so long as the assumptions used are adequately satisfied in practice.[5] Fig. 2 shows a simple non-linear circuit consisting of the series combination of a generator, impedance, Z, and a non-linear element. If the non-linear element can be adequately described by the power series of (6), then the solution is the same as given above. Equation 2 is obviously the same and therefore, so long as the assumptions of (10), etc., are satisfied, (31), (32), (33) and (34) represent accurate expressions of the currents.

BALANCED PUSH-PULL AMPLIFIER

A balanced push-pull amplifier is not often thought of as a feedback structure. However, Fig. 3 shows such a circuit having a cathode feedback impedance Z, common to both sides. The cathode impedance is usually used for bias and sometimes to assist in balancing. This circuit has been analyzed to determine the effect of second order distortions which are fed back via the cathode impedance, Z, even though they do not appear in the load. Note that for the perfectly balanced case odd order components cancel and even orders add across the cathode impedance. Thus, one might think of this as a structure with feedback only at even order distortion components.

Proceeding in the same fashion as for the previous example and assuming perfect balance we write the power series for the plate currents

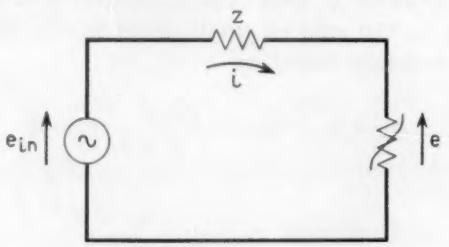


Fig. 2 — Non-linear circuit.

in terms of the grid-cathode voltages as

$$i_a = a_1 e_a + a_2 e_a^2 + a_3 e_a^3 \tag{38a}$$

$$i_b = a_1 e_b + a_2 e_b^2 + a_3 e_b^3 \tag{38b}$$

The input loops have the voltage relationships

$$e_a = A \cos \alpha t - Z(i_a + i_b) \tag{39a}$$

$$e_b = -A \cos \alpha t - Z(i_a + i_b) \tag{39b}$$

Finally, the grid-cathode voltages are expressed in Fourier series form.

$$e_a = \Sigma k_n \cos n\alpha t \tag{40a}$$

$$e_b = \Sigma j_n \cos n\alpha t \tag{40b}$$

Noting that the desired output is given by

$$\text{output current} = i_a - i_b \tag{41}$$

and observing that

$$k_1 \approx A \tag{42a}$$

$$j_1 \approx -A \tag{42b}$$

it can be shown that the dc output is

$$i_a|_{dc} - i_b|_{dc} = 0, \tag{43}$$

the fundamental is

$$i_a|_\alpha - i_b|_\alpha = \left(2a_1 A + 2A^2 \bar{A} \left[\frac{3a_3}{4} - \frac{2a_2^2 Z_0}{1 + 2a_1 Z_0} - \frac{a_2^2 Z_2}{1 + 2a_1 Z_2}\right]\right) \cos \alpha t \tag{44}$$

the second harmonic output current is

$$i_a|_{2\alpha} - i_b|_{2\alpha} = 0 \tag{45}$$

and the third harmonic output current is

$$i_a|_{3\alpha} - i_b|_{3\alpha} = A^3 \left[\frac{a_3}{2} - \frac{2a_2^2 Z_2}{1 + 2a_1 Z_2}\right] \cos 3\alpha t \tag{46}$$

The pattern here is the same as in the previous example, seconds are fed back to form thirds. Thus, while the balance removes second order

products from the output, nevertheless, the level of third order distortion can be materially increased. Even if the cathode impedance, Z, is bypassed, the dc component can produce gain changes. This is equally true of the ordinary single-sided cathode-biased amplifier, another "non-feedback" amplifier.

APPLICATIONS OF THE THEORY

During the development of the L3 coaxial system some work was done on a non-linear circuit to generate modulation products. The experimental results failed to check with existing theory by large factors and the theory described here was developed to explain the difference.

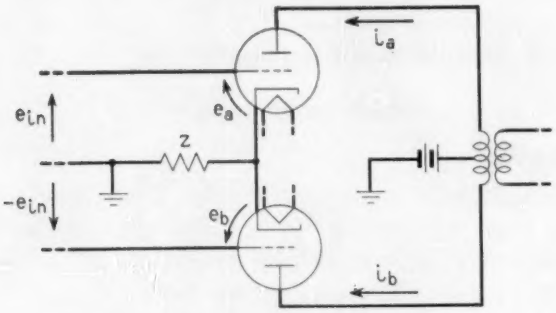


Fig. 3 — Balanced push-pull amplifier.

The deviations were then easily understood as, for example, a lack of feedback of direct current, a second-order product, yielding an expansion to third harmonic ratio appreciably different from three to one. With multiple frequency inputs the relative levels of other third order products were similarly affected. Application of the new theory which takes into account the magnitude and phase of the feedback yielded entirely satisfactory agreement.

The theory was later applied to various modulation characteristics of the L3 coaxial-system line amplifier.[3] In one case an effect was predicted by the theory which had not, at that time, been observed experimentally. In the L3 line amplifier the vacuum tubes are operated with local dc feedback from large cathode bias resistors suitably bypassed. The normal feedback loop does not extend to dc. The theory predicted that the compression would be reduced if this cathode dc feedback were removed. Typical values for the power series coefficients of the output

stage (437A tubes) are

$$a_1 = 0.045$$

$$a_2 = 0.014$$

$$a_3 = -0.0057$$

Using equation 37 and the approximation of large second order feedback, ($a_1Z_0 \gg 1$, $a_1Z_2 \gg 1$), the expansion is found to be

$$\text{expansion} = 1 - 0.240 \left(\frac{A}{1 + a_1Z_1}\right)\left(\overline{\frac{A}{1 + a_1Z_1}}\right)$$

When the dc feedback is removed leaving the second harmonic value unchanged, one calculates

$$\text{expansion} = 1 - 0.1433 \left(\frac{A}{1 + a_1Z_1}\right)\left(\overline{\frac{A}{1 + a_1Z_1}}\right)$$

The ratio of the calculated compression (negative expansion) increment is

$$\frac{0.240}{0.1433} = 1.673 = 4.4 \text{ db calculated}$$

Thus the calculation indicates that removal of the dc feedback would reduce the compression produced by the amplifier by 4.4 db.

The incremental loss in gain of the amplifier was measured with both normal operation and with the local dc feedback removed. The removal of the dc feedback reduced the measured compression increment by the ratio.

$$\frac{0.0945}{0.058} = 1.63 = 4.2 \text{ db measured}$$

Thus the predicted effect was verified and the amount confirmed. It might be noted that the dc feedback was retained in the design in spite of the somewhat greater compression because of its great value in stabilizing the current and transconductance of the tubes against aging effects.

A third application of the theory has been in the design of the ambient temperature compensation oscillator used in the pilot regulators of the L3 coaxial system.[4] Here it was necessary to obtain compression in an amplifier whose tube had a positive a_3 and tended to expand. Gain could not be expended on signal frequency feedback but, by the use of 20 db of dc feedback, compression from second order dc feedback was made greater than the expansion effect. Thus the amplifier compressed (lost

gain) as the signal increased. This effect, while small, nevertheless solved a serious motorboating problem. It also points up a technique by which gain changes with level can be balanced out. If a tube having a positive a_3 is provided with an appropriate amount of second order feedback (which always produces compression), then the two effects can be cancelled. This yields a gain which, to a good approximation, is independent of signal level. This is quite different from the technique of using a 90° feedback to convert "μ" gain changes into phase changes. Here there need be no gain or phase changes produced and feedback at the signal frequencies is unnecessary. However, the balance must be adjusted for the particular tube's modulation coefficients.

CONCLUSIONS

The analysis of cases with more than one input frequency are treated in the Appendix. Formulas are derived for second and third order products for up to three input frequencies. Certain of these results have also been expressed in terms of modulated tones to determine gain changes of sidebands relative to the carrier, etc. Comments on these results are included below.

Examining (31) to (34) pertaining to the output current for a single input frequency, it is interesting to note that, within the approximations, second order output is not affected by the feedback at other product frequencies. Of course, if the second order distortion, (31 and 33), were not large compared to the thirds this would no longer be the case. Third order outputs, (32), (34), show that the relative contributions of various seconds in forming thirds is not equal. For expansion the dc effect is normally twice the second harmonic effect while for third harmonic only the second harmonic contributes. Thus, the ratio of third harmonic to expansion, normally thought of as $\frac{1}{3}$, can vary widely depending on the relative feedbacks at dc and the second harmonic.

In general, it can be stated that the level of a given third order product is not an accurate indication of other third order products nor is it always a good indication of the same product at another set of frequencies.

In carrier systems products such as $2\alpha - \beta$ and $\alpha + \beta - \gamma$ tend to add by voltage when the fundamentals (and their products) are closely spaced because insufficient phase distortion occurs to break up the in-phase addition among amplifiers. Thus, products of this type often tend to dominate the linearity problem. These products involve frequency differences $\alpha - \beta$ and $\beta - \gamma$ for example, which fall at low frequencies where the second order feedback effects should be subject to control. Thus, by appropriate use of phase shifts or small feedback or even small positive feed-

back at these low (outside the band) frequencies it may, in some cases, be possible to either reduce the level of these products or break up their tendency to add by voltage.

In line with the above comment it should be noted that any tendency towards instability of the feedback loop can lead to abnormally high distortion levels. A second order interaction frequency may suffer from a $Z/(1 + a_1Z)$ factor in excess of $1/a_1$. Typically Z has an angle of 150 degrees or so in the cutoff region of a feedback loop. For large Z the factor is $1/a_1$ and independent of Z. However, where a_1Z drops to, say, $2/\sqrt{3}$ the magnitude of $Z/(1 + a_1Z)$ becomes $2/a_1$ at an angle of 60 degrees. Whether this doubles the interaction effect depends, of course, on the phase of the other terms but, in many cases, the effect will be appreciable. This example was for a feedback loop of conservative design. Where $1 + a_1Z$ approaches zero more closely the effect will be larger. Thus, an amplifier having poor stability margins may exhibit unusual modulation behavior for third order products involving second order differences falling in the frequency range of the poor stability margin.

Even in cases where the amplifier is not considered as having feedback, second order products may be returned to the input at frequencies outside the useful range to affect in-band products. This effect is perhaps most obvious in the dc case where expansion or compression is most likely to be affected. Such a dc feedback can, for example, cause the amplification of a short pulse to differ from that of a steady tone. Alternatively, a suddenly applied tone may, at first, produce one output level and then, following a transient dependent upon the cutoff of the dc feedback, settle down at a different level. Such effects are difficult to predict rigorously since they involve essentially a large number of input frequencies. However, the mechanisms involved are easily understood.

In the case of the balanced push-pull amplifier and similar circuits, second order distortion products can be fed back via stray paths without feedback of fundamentals. This can produce significant increases in third order distortion compared to a single-sided amplifier. This is analogous to the well-known fact that putting feedback around an amplifier rarely reduces third order products by the amount of the feedback. The reason is the same, the circuit change allows seconds to feed back and make thirds.

ACKNOWLEDGEMENT

The author wishes to acknowledge the assistance of E. G. Morton and E. F. O'Neill in obtaining experimental data.

BIBLIOGRAPHY

1. Llewellyn, F. B., Operation of Thermionic Vacuum Tube Circuits, B.S.T.J., **5**, pp. 433–462, July, 1926.
2. Llewellyn, F. B., Constant Frequency Oscillators, B.S.T.J., **11**, pp. 67–100, Jan., 1932.
3. Morris, L. H., Lovell, G. H., and Dickinson, F. R., The L3 Coaxial System — Amplifiers, B.S.T.J., **32**, pp. 879–914, July 1953.
4. Ketchledge, R. W. and Finch, T. R., The L3 Coaxial System — Equalization and Regulation, B.S.T.J., **32**, pp. 833–878, July 1953.
5. Gillies, A. W., The Application of Power Series to the Solution of Non-Linear Circuit Problems, Proc. I.E.E., **96**, No. 44, Part III, Nov., 1949.

APPENDIX

THREE FREQUENCY INPUT

In the case of three input frequencies the derivation of particular products is somewhat more complicated than in the case of a single frequency input and the number of products of interest tends to be greater. However, the methods and the assumptions are essentially the same.

The input signal is assumed to be of the form

$$e_{in} = A \cos \alpha t + B \cos \beta t + C \cos \gamma t \qquad \alpha > \beta > \gamma \tag{A1}$$

and the power series for the tube is still

$$i = a_1 e + a_2 e^2 + a_3 e^3 \tag{A2}$$

The loop equation (Fig. 1) is unchanged,

$$e = e_{in} - iZ \tag{A3}$$

The Fourier series for the grid-cathode voltage is taken as

$$e = \sum_{n\,pq} k_{n,p,q} \cos (n\alpha + p\beta + q\gamma), \qquad \alpha > \beta > \gamma \tag{A4}$$

which may be written as

$$\begin{aligned} e = {} & k_{0,0,0} + k_{1,0,0} \cos \alpha t + k_{0,1,0} \cos \beta t + k_{0,0,1} \cos \gamma t \\ & + k_{2,0,0} \cos 2\alpha t + k_{0,2,0} \cos 2\beta t + k_{0,0,2} \cos 2\gamma t \\ & + k_{1,-1,0} \cos (\alpha - \beta)t + k_{1,0,-1} \cos (\alpha - \gamma)t + k_{0,1,-1} \cos (\beta - \gamma)t \\ & + k_{1,1,0} \cos (\alpha + \beta)t + k_{1,0,1} \cos (\alpha + \gamma)t + k_{0,1,1} \cos (\beta + \gamma)t \\ & + k_{3,0,0} \cos 3\alpha t + k_{0,3,0} \cos 3\beta t + k_{0,0,3} \cos 3\gamma t \\ & + k_{1,1,1} \cos (\alpha + \beta + \gamma)t + k_{1,1,-1} \cos (\alpha + \beta - \gamma)t + \text{etc., etc., etc.} \end{aligned} \tag{A5}$$

In the above Fourier series for e, the dc through third order products total 32 terms. Consequently, the formation of e^2 involves 32^2 or approximately 1,000 multiplications. To form e^3 requires 32^3 or approximately

30,000 multiplications. Since the labor involved is obviously excessive the technique used is to select only the dominant terms in forming the desired products. In particular, the assumption that fundamentals are large compared to second order products and that these, in turn, are large compared to third order products is used repeatedly.

A typical third order product for a three frequency input is the one having the frequency $\alpha + \beta - \gamma$. In order to find the amplitude of this product the dominant terms in both e^2 and e^3 are selected by inspection. In e^2 the frequency, $\alpha + \beta - \gamma$, can be formed by a large variety of combinations. Some of these are

γ	and	$\alpha + \beta$	$\alpha + \gamma$	and	$2\alpha + \beta$
$\alpha + \beta$	and	γ	dc	and	$\alpha + \beta - \gamma$
$\beta - \gamma$	and	α	$\alpha + \beta - \gamma$	and	dc
α	and	$\beta - \gamma$	$2\alpha - \gamma$	and	$\alpha - \beta$
β	and	$\alpha - \gamma$	$\alpha - \beta$	and	$2\alpha - \gamma$
$\alpha - \gamma$	and	β		etc.	
$2\alpha + \beta$	and	$\alpha + \gamma$			

Recognizing that the dominant products are fundamentals $\times$ seconds we can write the e^2 terms of frequency $\alpha + \beta - \gamma$ as follows:

$$\begin{aligned} e^2\,|_{\alpha+\beta-\gamma} = {} & 2k_{1,1,0}k_{0,0,1}\cos(\alpha+\beta)t\cos\gamma t \\ & + 2k_{0,1,-1}k_{1,0,0}\cos(\beta-\gamma)t\cos\alpha t \\ & + 2k_{1,0,-1}k_{0,1,0}\cos(\alpha-\gamma)t\cos\beta t \end{aligned} \tag{A6}$$

When the indicated multiplications are carried out (A6) simplifies to the form given below. Note that the conjugate is used whenever its corresponding frequency subtracts in the formation of the desired product.

$$e^2\,|_{\alpha+\beta-\gamma} = [k_{1,1,0}\overline{k_{0,0,1}} + k_{0,1,-1}k_{1,0,0} + k_{1,0,-1}k_{0,1,0}]\cos(\alpha+\beta-\gamma)t \tag{A7}$$

In like manner the e^3 terms of frequency $\alpha + \beta - \gamma$ are observed to be dominated by fundamentals $\times$ fundamentals $\times$ fundamentals. Namely,

α	β	γ
α	γ	β
β	α	γ
β	γ	α
γ	α	β
γ	β	α

Thus,

$$e^3\,|_{\alpha+\beta-\gamma} = \tfrac{3}{4}k_{1,0,0}k_{0,1,0}\overline{k_{0,0,1}}\cos(\alpha+\beta-\gamma)t \tag{A8}$$

The coefficients for the fundamentals themselves are easily approximated in the same manner as was used to arrive at (26). However, it is still necessary to find the second order coefficients of e used in forming e^2 in (A7). These second order products are dominated by fundamentals x fundamentals and can therefore easily be shown to be

$$k_{1,1,0} = \frac{-a_2Z_{\alpha+\beta}}{1+a_1Z_{\alpha+\beta}}\,k_{1,0,0}k_{0,1,0} \tag{A9}$$

$$k_{0,1,-1} = \frac{-a_2Z_{\beta-\gamma}}{1+a_1Z_{\beta-\gamma}}\,k_{0,1,0}\overline{k_{0,0,1}} \tag{A10}$$

$$k_{1,0,-1} = \frac{-a_2Z_{\alpha-\gamma}}{1+a_1Z_{\alpha-\gamma}}\,k_{1,0,0}\overline{k_{0,0,1}} \tag{A11}$$

Since there is no original input signal of frequency $\alpha+\beta-\gamma$, (A3) can be written as

$$e = -Z_{\alpha+\beta-\gamma}[a_1e + a_2e^2 + a_3e^3]_{\alpha+\beta-\gamma} \tag{A12}$$

The subscript on both the impedance and the power series should be interpreted respectively as the value at this particular frequency and the content of this particular frequency. Inserting the appropriate values for e, e^2 and e^3 of frequency $\alpha+\beta-\gamma$ in (A12) one obtains,

$$\begin{aligned} k_{1,1,-1}(1+a_1Z_{\alpha+\beta-\gamma}) = -Z_{\alpha+\beta-\gamma}\Bigg[&\frac{-a_2^{\,2}}{1+a_1Z_{\alpha+\beta}}\,k_{1,0,0}k_{0,1,0}\overline{k_{0,0,1}} \\ &+\frac{-a_2^{\,2}}{1+a_1Z_{\beta-\gamma}}\,k_{0,1,0}\overline{k_{0,0,1}}k_{1,0,0} + \frac{-a_2^{\,2}}{1+a_1Z_{\alpha-\gamma}}\,k_{1\,0,0}\overline{k_{0,0,1}}k_{0,1,0} \\ &+a_3\frac{6}{4}\,k_{1,0,0}k_{0,1,0}\overline{k_{0,0,1}}\Bigg] \end{aligned} \tag{A13}$$

Note that

$$i_{\alpha+\beta-\gamma} = \frac{-k_{1,1,-1}}{Z_{\alpha+\beta-\gamma}} \tag{A14}$$

Thus,

$$\begin{aligned} i_{\alpha+\beta-\gamma} = \frac{1}{1+a_1Z_{\alpha+\beta-\gamma}}\Bigg[&\frac{3a_3}{2} - \frac{a_2^{\,2}Z_{\alpha+\beta}}{1+a_1Z_{\alpha+\beta}} - \frac{a_2^{\,2}Z_{\beta-\gamma}}{1+a_1Z_{\beta-\gamma}} \\ &- \frac{a_2^{\,2}Z_{\alpha-\gamma}}{1+a_1Z_{\alpha-\gamma}}\Bigg]k_{1,0,0}\overline{k_{0,0,1}}k_{0,1,0}\cos(\alpha+\beta-\gamma)t \end{aligned} \tag{A15}$$

Approximating the fundamentals by

$$k_{1,0,0} = \frac{A}{1 + a_1 Z_\alpha} \tag{A16}$$

$$k_{0,1,0} = \frac{B}{1 + a_1 Z_\beta} \tag{A17}$$

$$k_{0,0,1} = \frac{C}{1 + a_1 Z_\gamma} \tag{A18}$$

We obtain for the desired product

$$i_{\alpha+\beta-\gamma} = \frac{1}{1 + a_1 Z_{\alpha+\beta-\gamma}} \left[\frac{3a_3}{2} - \frac{a_2^2 Z_{\alpha+\beta}}{1 + a_1 Z_{\alpha+\beta}} - \frac{a_2^2 Z_{\beta-\gamma}}{1 + a_1 Z_{\beta-\gamma}} - \frac{a_2^2 Z_{\alpha-\gamma}}{1 + a_1 Z_{\alpha-\gamma}} \right] \left(\frac{A}{1 + a_1 Z_\alpha} \right) \left(\frac{B}{1 + a_1 Z_\beta} \right) \left(\frac{\overline{C}}{1 + a_1 Z_\gamma} \right) \cos(\alpha + \beta - \gamma)t \tag{A19}$$

A similar derivation for the $\alpha + \beta + \gamma$ product yields the same expression except that the conjugate is removed and $+\gamma$ replaces $-\gamma$ on the impedance and frequency subscripts wherever $-\gamma$ appears above. Thus,

$$i_{\alpha+\beta+\gamma} = \frac{1}{1 + a_1 Z_{\alpha+\beta+\gamma}} \left[\frac{3a_3}{2} - \frac{a_2^2 Z_{\alpha+\beta}}{1 + a_1 Z_{\alpha+\beta}} - \frac{a_2^2 Z_{\beta+\gamma}}{1 + a_1 Z_{\beta+\gamma}} - \frac{a_2^2 Z_{\alpha+\gamma}}{1 + a_1 Z_{\alpha+\gamma}} \right] \left(\frac{A}{1 + a_1 Z_\alpha} \right) \left(\frac{B}{1 + a_1 Z_\beta} \right) \left(\frac{C}{1 + a_1 Z_\gamma} \right) \cos(\alpha + \beta + \gamma)t \tag{A20}$$

Another product of interest for three frequency inputs is expansion. As a typical case, the expansion of β in the presence of α, β, and γ is derived below. The method used relies on the same assumptions. Since the distortion product has the frequency β, in e^2, the following fundamentals x seconds terms dominate.

α	and	$\alpha - \beta$	dc	and	β
$\alpha - \beta$	and	α	γ	and	$\beta - \gamma$
$\alpha + \beta$	and	α	$\beta - \gamma$	and	γ
2β	and	β	γ	and	$\beta + \gamma$
β	and	2β	$\beta + \gamma$	and	γ
β	and	dc			

This simplifies to

$$e^2 |_\beta = [2k_{0,0,0}k_{0,1,0} + k_{0,2,0}\overline{k_{0,1,0}} + k_{1\,1,0}\overline{k_{1,0,0}} + k_{1,0,0}\overline{k_{1,-1,0}} + k_{0,1,-1}k_{0,0,1} + k_{0,1,1}\overline{k_{0,0,1}}] \cos \beta t \tag{A21}$$

In e^3 fundamentals $\times$ fundamentals $\times$ fundamentals dominate as follows,

α	β	α	3 terms
γ	β	γ	3 terms
β	β	β	1 term

This yields

$$e^3 |_\beta = \tfrac{3}{2}k_{1.0.0}\overline{k_{1,0,0}}k_{0,1,0} + \tfrac{3}{2}k_{0,0.1}\overline{k_{0,0,1}}k_{0,1,0} + \tfrac{3}{4}k_{0}^{2}{}_{,1,0}\overline{k_{0,1,0}} \tag{A22}$$

To proceed further again requires the determination of a number of second order products appearing in e^2. Note that these are the products which when fed back will beat with fundamentals to form the desired third order product. Since these second order products are dominated again by fundamentals $\times$ fundamentals they are easily shown to be

$$k_{0,0,0} = \frac{-a_2Z_0}{2(1 + a_1Z_0)} (k_{1,0,0}\overline{k_{1,0,0}} + k_{0,1,0}\overline{k_{0,1,0}} + k_{0,0\,1}\overline{k_{0,0,1}}) \tag{A23}$$

$$k_{0,2,0} = \frac{-a_2Z_{2\beta}}{2(1 + a_1Z_{2\beta})} (k_{0\,1,0})^2 \tag{A24}$$

$$k_{1,1,0} = \frac{-a_2Z_{\alpha+\beta}}{(1 + a_1Z_{\alpha+\beta})} k_{1,0,0}k_{0,1,0} \tag{A25}$$

$$k_{1,-1,0} = \frac{-a_2Z_{\alpha-\beta}}{1 + a_1Z_{\alpha-\beta}} k_{1,0,0}\overline{k_{0,1,0}} \tag{A26}$$

$$k_{0,1,-1} = \frac{-a_2Z_{\beta-\gamma}}{1 + a_1Z_{\beta-\gamma}} k_{0,1,0}\overline{k_{0,0,1}} \tag{A27}$$

$$k_{0,1,1} = \frac{-a_2Z_{\beta+\gamma}}{1 + a_1Z_{\beta+\gamma}} k_{0,1,0}k_{0,0,1} \tag{A28}$$

The next step is to rewrite (A3) as

$$k_{0,1,0} = B - Z_\beta(a_1k_{0,1,0} + a_2e^2 |_\beta + a_3e^3 |_\beta) \tag{A29}$$

which reduces to

$$k_{0,1,0} = \frac{B}{1 + a_1Z_\beta} - \frac{Z_\beta}{1 + a_1Z_\beta} (a_2e^2 |_\beta + a_3e^3 |_\beta) \tag{A30}$$

Substituting the terms of e^2 and e^3 of frequency β we obtain,

$$k_{0,1,0} = \frac{B}{1 + a_1Z_\beta} - \frac{Z_\beta}{1 + a_1Z_\beta}\Bigg[(a_2k_{0,1,0})\left(\frac{-a_2Z_0}{1 + a_1Z_0}(k_{1,0,0}\overline{k_{1,0,0}} + k_{0,1,0}\overline{k_{0,1,0}} + k_{0,0,1}\overline{k_{0,0,1}}) - \frac{a_2Z_{2\beta}}{2(1 + a_1Z_{2\beta})}k_{0,1,0}\overline{k_{0,1,0}} - \frac{a_2Z_{\alpha+\beta}}{1 + a_1Z_{\alpha+\beta}}k_{1,0,0}\overline{k_{1,0,0}} - \frac{a_2Z_{\alpha-\beta}}{1 + a_1Z_{\alpha-\beta}}k_{1,0,0}\overline{k_{1,0,0}} - \frac{a_2Z_{\beta-\gamma}}{1 - a_1Z_{\beta-\gamma}}k_{0,0,1}\overline{k_{0,0,1}} - \frac{a_2Z_{\beta+\gamma}}{1 + a_1Z_{\beta+\gamma}}k_{0,0,1}\overline{k_{0,0,1}}\right) + a_3k_{0,1,0}(\tfrac{3}{4}k_{0,1,0}\overline{k_{0,1,0}} + \tfrac{3}{2}k_{1,0,0}\overline{k_{1,0,0}} + \tfrac{3}{2}k_{0,0,1}\overline{k_{0,0,1}})\Bigg] \tag{A31}$$

Making use of values of the fundamentals as given in (16), (17), (18) and the relationship

$$i_\beta = -\frac{k_{0,1,0} - B}{Z_\beta} \tag{A32}$$

we find for the output current of frequency β

$$i_\beta = \frac{a_1B}{1 + a_1Z_\beta}\Bigg[1 + \frac{1}{1 + a_1Z_\beta}\left(\frac{B}{1 + a_1Z_\beta}\right)\left(\overline{\frac{B}{1 + a_1Z_\beta}}\right)\left(\frac{3a_3}{4a_1} - \frac{a_2^2Z_0}{a_1(1 + a_1Z_0)} - \frac{a_2^2Z_{2\beta}}{2a_1(1 + a_1Z_{2\beta})}\right) + \frac{1}{1 + a_1Z_\beta}\left(\frac{A}{1 + a_1Z_\alpha}\right)\left(\overline{\frac{A}{1 + a_1Z_\alpha}}\right)\left(\frac{3a_3}{2a_1} - \frac{a_2^2Z_0}{a_1(1 + a_1Z_0)} - \frac{a_2^2Z_{\alpha+\beta}}{a_1(1 + a_1Z_{\alpha+\beta})} - \frac{a_2^2Z_{\alpha-\beta}}{a_1(1 + a_1Z_{\alpha-\beta})}\right) + \frac{1}{1 + a_1Z_\beta}\left(\frac{C}{1 + a_1Z_\gamma}\right)\left(\overline{\frac{C}{1 + a_1Z_\gamma}}\right)\left(\frac{3a_3}{2a_1} - \frac{a_2^2Z_0}{a_1(1 + a_1Z_0)} - \frac{a_2^2Z_{\beta-\gamma}}{a_1(1 + a_1Z_{\beta-\gamma})} - \frac{a_2^2Z_{\beta+\gamma}}{a_1(1 + a_1Z_{\beta+\gamma})}\right)\Bigg]\cos\beta t \tag{A33}$$

The gain expansion is readily obtained from the above using the relationship

$$\text{gain expansion} = \frac{i_\beta}{\dfrac{a_1B}{1 + a_1Z_\beta}\cos\beta t} \tag{A34}$$

It is interesting to note that the influences of the other two fundamentals involve feedbacks at sum and difference frequencies relative to

β whereas the corresponding term for β involves the second harmonic and dc feedback. The latter also appears with the other fundamentals.

MODULATED CARRIER EXPANSION

The preceding analysis of three frequency inputs can also be applied to the situation where the signal consists of an ordinary amplitude modulated carrier. Such a modulated carrier consists of a carrier frequency with two sideband frequencies, one above and one below the carrier. To simplify the results given below it is assumed that the feedback is the same for the sidebands as it is for the carrier. The index of modulation is expressed as m and the modulating frequency is expressed as Δ. Thus, in the notation of the preceding section one has the relationships

$$A = C = \frac{m}{2} B \tag{A35}$$

$$\alpha = \beta + \Delta \tag{A36}$$

$$\gamma = \beta - \Delta \tag{A37}$$

$$Z_{\beta+\Delta} = Z_\beta = Z_{\beta-\Delta}$$

It is further assumed that the feedback is unchanged between the second harmonics of the sidebands and carrier so that

$$Z_{2(\beta+\Delta)} = Z_{2\beta} = Z_{2(\beta-\Delta)} \tag{A38}$$

Taking (A33) and equivalent expressions for i_α and i_γ it may be shown that

gain expansion of the carrier = 1

$$+ \frac{1}{1 + a_1 Z_\beta} \left(\frac{B}{1 + a_1 Z_\beta} \right) \left(\overline{\frac{B}{1 + a_1 Z_\beta}} \right) \left[\frac{3a_3}{4a_1} (1 + m^2) \right.$$
$$- \frac{a_2^2 Z_0}{a_1(1 + a_1 Z_0)} \left(1 + \frac{m^2}{2} \right) - \frac{a_2^2 Z_{2\beta}}{a_1(1 + a_1 Z_{2\beta})} \left(\frac{1 + m^2}{2} \right) \tag{A39}$$
$$\left. - \frac{a_2^2 Z_\Delta}{a_1(1 + a_1 Z_\Delta)} \left(\frac{m^2}{2} \right) \right]$$

gain expansion of either sideband = 1

$$+ \frac{1}{1 + a_1 Z_\beta} \left(\frac{B}{1 + a_1 Z_\beta} \right) \left(\overline{\frac{B}{1 + a_1 Z_\beta}} \right) \left[\frac{3a_3}{2a_1} \left(1 + \frac{3}{8} m^2 \right) \right.$$
$$- \frac{a_2^2 Z_0}{a_1(1 + a_1 Z_0)} \left(1 + \frac{m^2}{2} \right) - \frac{a_2^2 Z_{2\beta}}{a_1(1 + a_1 Z_{2\beta})} \left(1 + \frac{3m^2}{8} \right) \tag{A40}$$
$$\left. - \frac{a_2^2 Z_\Delta}{a_1(1 + a_1 Z_\Delta)} (1) - \frac{a_2^2 Z_{2\Delta}}{a_1(1 + a_1 Z_{2\Delta})} \left(\frac{m^2}{4} \right) \right]$$

and the expansion of the modulation index is

$$\text{gain expansion of sideband relative to carrier} = 1 + \frac{1}{1 + a_1 Z_\beta}\left(\frac{B}{1 + a_1 Z_\beta}\right)\left(\frac{\overline{B}}{1 + a_1 Z_\beta}\right)\left[\frac{3a_3}{4a_1}\left(1 - \frac{m^2}{4}\right) - \frac{a_2^2 Z_{2\beta}}{a_1(1 + a_1 Z_{2\beta})}\left(\frac{1}{2} - \frac{m^2}{8}\right) - \frac{a_2^2 Z_{2\Delta}}{a_1(1 + a_1 Z_\Delta)}\left(1 - \frac{m^2}{2}\right) - \frac{a_2^2 Z_{2\Delta}}{a_1(1 + a_1 Z_{2\Delta})}\left(\frac{m^2}{4}\right)\right] \tag{A41}$$

Note that the dc term drops out of the expansion of the sidebands relative to the carrier because the dc term affects them equally.

TWO FREQUENCY INPUT

The third order products, $2\alpha \pm \beta$, are often of importance in carrier systems. They cannot be found simply by substitution of variables in (A19) or (A20) since, for example, α, β, γ, can be formed in six ways and α, α, β, in only three. Thus, the a_3 term has a coefficient of $\frac{3}{4}$ instead of $\frac{6}{4}$. This is intended to sound a note of caution in merely changing variables to find other products.

Carrying through the complete calculations for $2\alpha \pm \beta$ one obtains,

$$i_{2\alpha-\beta} = \frac{1}{1 + a_1 Z_{2\alpha-\beta}}\left(\frac{A}{1 + a_1 Z_\alpha}\right)^2\left(\frac{\overline{B}}{1 + a_1 Z_\beta}\right)\left[\frac{3c_3}{4} - \frac{a_2^2 Z_{2\alpha}}{2(1 + a_1 Z_{2\alpha})} - \frac{a_2^2 Z_{\alpha-\beta}}{1 + a_1 Z_{\alpha-\beta}}\right]\cos(2\alpha - \beta)t \tag{A42}$$

and

$$i_{2\alpha+\beta} = \frac{1}{1 + a_1 Z_{2\alpha+\beta}}\left(\frac{A}{1 + a_1 Z_\alpha}\right)^2\left(\frac{B}{1 + a_1 Z_\beta}\right)\left[\frac{3a_3}{4} - \frac{1}{2}\frac{a_2^2 Z_{2\alpha}}{(1 + a_1 Z_{2\alpha})} - \frac{a_2^2 Z_{\alpha+\beta}}{1 + a_1 Z_{\alpha+\beta}}\right]\cos(2\alpha + \beta)t \tag{A43}$$

Analysis of Switching Networks

By C. Y. LEE

(Manuscript received August 2, 1955)

Using a simplified model, an analysis of switching networks is presented. Methods for finding characteristics of a network such as blocking probability, retrial and connection-time distributions are given. The problem of equivalent crosspoint minimization is also considered.

1. INTRODUCTION

The application of probability theory to telephone traffic problems owes its origin to the pioneering work of A. K. Erlang, T. C. Fry, E. C. Molina and others.[1, 2, 3] Since then much has been written on these and other related problems. On the other hand, except for several recent papers on the subject,[4, 5, 6, 7] the literature on the application of probability theory to large size switching systems has been comparatively meager, mainly because the sheer size and complexity of these systems tend to render exact analysis unmanageable.

To fix ideas, let us peer into a telephone central office and point our attention at a single link (or crosspoint) in the system. As time progresses the link becomes busy and idle in some fashion and gives rise to a sequence of pairs of observations:

$$(t_0\text{, busy}), \qquad (t_1\text{, change to idle}), \qquad (t_2\text{, change to busy}), \cdots$$

Let x_t be a function such that x_t is 1 if the link is idle and is 0 if the link is busy, then the sequence of pairs of observations corresponds to the behavior of x_t as t changes. A plot of the values of x_t versus t would look perhaps as shown in Fig. 1.1. The function x_t (or the sequence of pairs of observations) is one of a large family of possible functions for the link. Since there are in general several thousand such links in a telephone central office, a complete description would involve several thousand families of functions x_t. We may add that the situation is made somewhat worse by the fact that these links are not independent of each other, for example, the establishment of a telephone conversation involves in general not one but several links in series.

In order to derive useful results, we shall consider a simplified model in which the links are assumed to be independent and we shall further restrict the families of functions x_t to only those which obey certain rules. The mathematical model used for a switching network is a probability linear-graph introduced in this paper. The basic steps we follow in the analysis of switching networks consist of (1) representing a switching network by a probability linear-graph and (2) using the graph as a probability model to calculate all characteristics of interest of the switching network. Once step (1) is effected, step (2) falls into the domain of probability theory in which standard techniques are available.

In Section 2 of this paper, the more basic aspects of switching network analysis are considered and properties of probability linear-graphs are discussed. Those who are mainly interested in numerical results may begin directly with Section 3 where representation of switching networks by graphs, probability of blocking of various networks and the problem of equivalent crosspoint minimization are presented. Using the graph model, methods for calculating average retrial time for blocked calls and network blocking due to excessive time required in the breakdown of crosspoints in a gas tube network are then given in Section 4.

2. PROBABILITY LINEAR-GRAPHS

It would be appropriate for us to associate with each link a class of binary-valued functions ω such that $\omega(t) = 0$ or 1 (i.e. the link is either busy or idle) at any given time t. It is more convenient, however, for us to associate with each link a random variable x_t (Reference 8, Chapter 17) such that at any given time t, $x_t = 0$ or 1. Furthermore, we assume that the random variable x_t has a stationary probability distribution.*

Definition 2.1. A (two-terminal) probability linear-graph G (or simply a graph if no confusion arises) is a finite, oriented, connected, cycle-free

* More strictly, we let Ω be the space of all binary-valued functions ω such that $\omega(t) = 0$ or 1, $-\infty < t < \infty$. Let

$$\{x_t\ ,\ -\infty < t < \infty\}$$

be a stochastic process with discrete probability distributions

$$f_{t_1\ ,\ t_2\ ,\ \cdots\ t_n}(\delta_1\ ,\ \delta_2\ ,\ \cdots\ ,\ \delta_n) = Pr[x_{t_i} = \delta_i\ ,\ i = 1, 2, \cdots\ , n]$$

$$\delta_i = 0 \text{ or } 1,\ i = 1, 2, \cdots\ , n$$

where x_s is the s^{th} coordinate function of Ω; that is $x_s(\omega) = \omega(s)$. Furthermore, we assume that this process is strictly stationary; that is

$$f_{t_1\ ,\ t_2\ ,\ \cdots\ ,\ t_n}(\delta_1\ ,\ \delta_2\ ,\ \cdots\ ,\ \delta_n) = f_{t_1+h\ ,\ t_2+h\ ,\ \cdots\ ,\ t_n+h}(\delta_1\ ,\ \delta_2\ ,\ \cdots\ ,\ \delta_n)$$

For each fixed $\omega \in \Omega$, x_t is a binary-valued function whose domain is $-\infty < t < \infty$. We then follow Reference 9 and call x_t a random variable indexed by t with a stationary probability distribution. For the existence and consistency of these stochastic processes, see Reference 9, Chapter I.

linear-graph* with at least two nodes such that (i) there exists a pair of nodes, called respectively an originating and terminating node, of G; this assignment being determined initially, (ii) with each link† of G is associated a random variable $x_t^{(i)}$, the index i running over all links of G such that $x_t^{(i)}$ are mutually independent and have stationary probability distributions

$$f_{t_1, t_2, \cdots, t_n}^{(i)}(\delta_1, \delta_2, \cdots, \delta_n) = Pr[x_{t_j}^{(i)} = \delta_j, j = 1, 2, \cdots, n]$$

$$\delta_j = 0 \text{ or } 1, \qquad j = 1, 2, \cdots, n$$

In what follows, by a directed path of G is meant a directed path between the originating and terminating nodes of G. Moreover, whenever there is no confusion, the letter x will be used to denote the vector

$$(x_t^{(1)}, x_t^{(2)}, \cdots)$$

Given a probability linear-graph, the usual problem is to find the

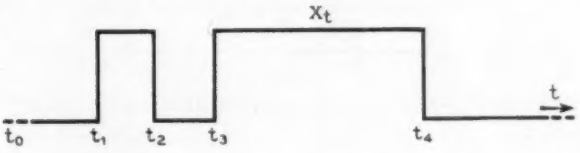


Fig. 1.1 — Observations of idle and busy conditions of a link.

probability distribution of some binary-valued function φ of independent random variables $x_t^{(i)}$. We begin this section by looking first at the simplest cases.

2.1 *Series-Parallel Graphs.*

Consider the series graph shown in Figure 2.1 with $N + 1$ nodes and N links. Let φ_t be the function given by

$$\varphi_t(x) = \prod_{i=1}^{N} x_t^{(i)}$$

Physically, if we interpret $x_t^{(i)} = 1$ as the event of link i being idle at time t, then $\varphi_t(x) = 1$ corresponds to the event that there is at least one directed path of G (in this case exactly one) idle at time t. In view of this interpretation, φ_t is called the *connection function* of G. The prob-

* A linear-graph is finite if both the sets of nodes and branches are non-empty and finite; it is connected if there exists at least one path (chain of branches) between each pair of nodes; it is oriented if the branches are all directed; it is cycle-free if there exists no *directed* path between any node and itself.

† Here, directed branches are called links.

abilities

$$P_\tau(1) = Pr[\varphi_\tau(x) = 1] \quad \text{and} \quad P_\tau(0) = Pr[\varphi_\tau(x) = 0] = 1 - P_\tau(1) \tag{2.1}$$

are called respectively the *linking* and *blocking* probabilities of G. Clearly,

$$\begin{aligned} P_\tau(1) &= Pr[\varphi_\tau(x) = 1] = Pr[x_\tau^{(i)} = 1, i = 1, 2, \cdots, N] \\ &= \prod_{i=1}^{N} Pr[x_\tau^{(i)} = 1] = \prod_{i=1}^{N} f_\tau^{(i)}(1) \end{aligned} \tag{2.2}$$

Next, let

$$P_{\tau,\,\tau+t}(\delta_1, \delta_2) = Pr[\varphi_\tau(x) = \delta_1, \varphi_{\tau+t}(x) = \delta_2] \tag{2.3}$$

then

$$\begin{aligned} P_{\tau,\tau+t}(1, 1) &= Pr[x_\tau^{(i)} = 1, x_{\tau+t}^{(i)} = 1; i = 1, 2, \cdots, N] \\ &= \prod_{i=1}^{N} f_{\tau,\tau+t}(1, 1) \end{aligned} \tag{2.4}$$

Since $x_t^{(i)}$ have stationary distributions, $P_\tau(\delta)$, $f_\tau^{(i)}(\delta)$ and $P_{\tau,\,\tau+t}(\delta_1, \delta_2)$

FIG. 2.1 — A series graph with N links.

are all independent of τ so that they may be written respectively $P(\delta)$, $f^{(i)}(\delta)$ and $P_t(\delta_1, \delta_2)$. Moreover,

$$P(1) = P_t(0, 1) + P_t(1, 1) = P_t(1, 0) + P_t(1, 1)$$

so that

$$\begin{aligned} P_t(0, 1) &= P_t(1, 0) = P(1) - P_t(1, 1) \\ &= \prod_{i=1}^{N} f^{(i)}(1) - \prod_{i=1}^{N} f_t^{(i)}(1, 1) \end{aligned} \tag{2.5}$$

Finally, let the means of $x_t^{(i)}$ and φ_t be denoted by

$$q^{(i)} = E(x_t^{(i)}), \qquad Q = E(\varphi_t) \tag{2.6}$$

Then, for a series graph, the following relations obtain:

$$Q = \prod_i q^{(i)} \tag{2.7}$$

$$P_t(1, 1) = \prod_i f_t^{(i)}(1, 1) \tag{2.8}$$

$$P_t(0, 1) = P_t(1, 0) = \prod_i q^{(i)} - \prod_i f_t^{(i)}(1, 1) \tag{2.9}$$

$$P_t(0, 0) = 1 - 2\prod_i q^{(i)} + \prod_i f_t^{(i)}(1, 1) \tag{2.10}$$

In the case of a parallel graph of N links (Figure 2.2), the connection function φ_t is given by*

$$\varphi_t(x) = 1 - \prod_i (1 - x_t^{(i)}).$$

Using the same notations as before, we obtain, for the parallel graph

$$Q = 1 - \prod_i (1 - q^{(i)}) \tag{2.11}$$

$$P_t(0, 0) = \prod_i f_t^{(i)}(0, 0) \tag{2.12}$$

$$P_t(0, 1) = P_t(1, 0) = \prod_i (1 - q^{(i)}) - \prod_i f_t^{(i)}(0, 0) \tag{2.13}$$

$$P_t(1, 1) = 1 - 2\prod_i (1 - q^{(i)}) + \prod_i f_t^{(i)}(0, 0) \tag{2.14}$$

2.2 *General Probability Linear-Graphs*

Let us first make the notion of a connection function of a graph more precise.

Let G be a probability linear-graph with N links. Let B be the set of all directed paths of G. Then each $\beta_i \in B$ is composed of series links of G with associated link random variables $x_t^{(i_1)}, x_t^{(i_2)}, \cdots$. To each $\beta_j \in B$, assign a new random variable $y_t^{(j)}$ as a function of $x_t^{(i_1)}, x_t^{(i_2)}, \cdots$ such that $y_t^{(j)} = 1$ if and only if $x_t^{(i_k)} = 1, k = 1, 2, \cdots$ and otherwise $y_t^{(j)} = 0$.

Definition 2.2. The graph G^* composed of all parallel links β, $\beta \in B$ with link random variables $y_t^{(1)}, y_t^{(2)}, \cdots$ is said to be the canonical form of G.

Definition 2.3. A binary-valued function φ_τ (itself a random variable) of the vector $y = (y_t^{(1)}, y_t^{(2)}, \cdots)$ such that $\varphi_\tau(y) = 0$ if and only if

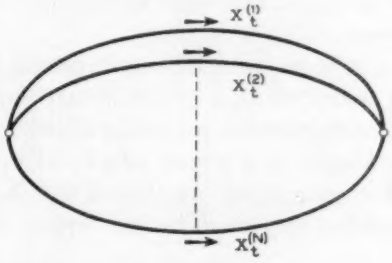


FIG. 2.2 — A parallel graph with N links.

* Note that by the change of variables $x_t'^{(i)} = 1 - x_t^{(i)}$ and $\varphi_t' = 1 - \varphi_t$, the situation here becomes identical with that for the series graphs.

$y = 0$ and $\varphi_\tau(y) = 1$ otherwise is said to be the connection function of G. The probabilities $Pr[\varphi_\tau(y) = 1]$ and $Pr[\varphi_\tau(y) = 0]$ are called respectively the *linking* and *blocking* probabilities of G. The conditional probability $Pr[\varphi_{\tau+t}(y) = 0 \mid \varphi_\tau(y) = 0]$ is called the *retrial distribution* of G.

We note that the physical interpretation of φ_τ is as it should be; $\varphi_\tau(y) = 1$ if there is at least one idle path through G and $\varphi_\tau(y) = 0$ if otherwise. The physical interpretations of linking and blocking probabilities and retrial distribution are self-evident.

We now let x denote the vector $(x_t^{(1)}, x_t^{(2)}, \cdots, x_t^{(N)})$ and let x_t^* denote the values of x at time t. At a given time t, let $J(\delta)$, $\delta = 0$ or 1, be the set of values of x which makes $\varphi_t(x) = \delta$. Since x, as a vector, has a stationary distribution, it follows that

$$\begin{aligned} Pr[\varphi_{t_1+h}(x) = \delta_1, \cdots, \varphi_{t_n+h}(x) = \delta_n] \\ &= Pr[x_{t_1+h}^* \in J(\delta_1), \cdots, x_{t_n+h}^* \in J(\delta_n)] \\ &= Pr[x_{t_1}^* \in J(\delta_1), \cdots, x_{t_n}^* \in J(\delta_n)] \\ &= Pr[\varphi_{t_1}(x) = \delta_1, \cdots, \varphi_{t_n}(x) = \delta_n] \end{aligned}$$

Thus, we have the elementary but important fact that

Theorem 2.1. The connection function φ_t for a probability linear-graph G has a stationary distribution.

An immediate consequence of Theorem 2.1 is that the linking (or blocking) probability of G is independent of the time t and is expressible as a polynomial with integral coefficients in terms of the linking probabilities $q^{(i)}$ of the links of G. To put if differently, Theorem 2.1 asserts that in order to compute the linking (or blocking) probability of G, it suffices for us to consider the process as one consisting of repeated simultaneous tossing of N skewed coins, a process in which time does not enter. It should be remarked that this fact has long been known to telephone traffic engineers.

Another remark is in order here although a precise statement would be needlessly long. In many cases, a graph G may contain several components G_j. If these components are well-defined, it is always possible to replace them by single links whose associated random variables are $\varphi_t^{(j)}$ where $\varphi_t^{(j)}$ is the connection function of G_j. Actual computations may be greatly simplified by a repeated application of this procedure.

2.3 *An Example of a Non-Stationary Process*

In Section 2.1 we had stipulated that the link random variables of a graph have stationary probability distributions. Let us now consider a

(somewhat arbitrary) example in which the stationary property is dropped.

Let ℓ_1 be a single link and let Ω be the family of functions ω associated with this link such that for the internal $0 \leqq \lambda < T$,

$$\omega(t) = 0; \qquad t < \lambda, \qquad t \geqq T$$

$$\omega(t) = 1; \qquad T > t \geqq \lambda$$

Thus, we may index functions of Ω by λ. A subset Ω_1 of Ω is said to be measurable if the corresponding indexing set Λ_1 is Borel measurable on the interval $[0, T)$ and in such cases, we identify the measure of Ω_1 with the Borel measure of Λ_1.

Let x_s be the s^{th} coordinate function of Ω; that is, $x_s(\omega) = \omega(s)$. Then,

$$Pr[x_t = 1] = \frac{t}{T}, \qquad 0 \leq t < T$$

and

$$Pr[x_t = 1] = 0 \quad \text{otherwise.}$$

Hence the probability of blocking P_1 for ℓ_1 is zero for $t < 0$, $t \geqq T$ and is t/T for $0 \leqq t < T$. That is, the blocking probability is itself a function of t.

If we restrict our attention to the time interval $[0, T)$, we may define the mean blocking $\overline{P}_1$ as the time average of P on $[0, T)$.

Then

$$\overline{P}_1 = \frac{1}{T}\int_0^T \frac{t}{T}\,dt = \tfrac{1}{2}$$

Suppose we now have two such links ℓ_1, ℓ_2, independent of each other, in parallel. Then the blocking probability of the parallel graph is

$$P = P_1^2 = \left(\frac{t}{T}\right)^2$$

The mean blocking for the parallel graph is

$$\overline{P} = \frac{1}{T}\int_0^T \left(\frac{t}{T}\right)^2 dt = \tfrac{1}{3}$$

Thus, it is erroneous here to say that the mean blocking of the parallel graph is the square of the mean blocking of each path.

Strictly speaking, the assumption of stationary distributions for link random variables is invalid in practice, since obviously a telephone cen-

tral office handles more calls at noon than at midnight. However, it is the busy-hour-traffic which concerns the telephone traffic engineers, so that the assumption is in general a reasonable one.

2.4 *Linking Probability*

In this section, only linking (and blocking) probabilities of a probability linear-graph will be considered. Other characteristics of a graph (retrial distribution, connection-time distribution etc.) will be studied in a later section.

Let G be a probability linear-graph and let the random variables associated with the links of G be $x^{(1)}, x^{(2)}, \cdots$ where these random variables are no longer functions of time. Let us view our probability process as one consisting of repeated trials (or, say, tosses of batches of coins) in which, for each trial, the probabilities* of success and failure for the links are respectively

$$q^{(i)} = Pr[x^{(i)} = 1], \qquad p^{(i)} = Pr[x^{(i)} = 0] = 1 - q^{(i)}, \quad (2.15)$$

$$i = 1, 2, \cdots$$

Let φ be the connection function for G and denote by Q and P

$$Q = Pr[\varphi(x) = 1], \qquad P = 1 - Q = Pr[\varphi(x) = 0] \quad (2.16)$$

where x is the vector $(x^{(1)}, x^{(2)}, \cdots)$. Then Q and P are respectively the linking and blocking probabilities of G.

In the case G is a series graph of N links, we have, from Section 2.1,

$$Q = \prod_{i=1}^{N} q^{(i)}, \qquad P = 1 - Q. \quad (2.17)$$

Similarly, when G is a parallel graph of N links,

$$Q = 1 - \prod_{i=1}^{N} (1 - q^{(i)}), \qquad P = 1 - Q. \quad (2.18)$$

Thus for any series-parallel graph G, a combination of (2.17) and (2.18) will yield the linking and blocking probabilities of G.

Example 2.1. Consider the series-parallel graph G shown in Fig. 2.3 where the link random variables are denoted by

$$x_1^{(1)}, \quad x_1^{(2)}, \cdots, \quad x_1^{(10)}; \quad x_2^{(1)}, \quad x_2^{(2)}, \cdots, \quad x_2^{(10)}$$

* $p^{(i)}$ is usually called the *occupancy* of link i.

and

$$x_3^{(1)}, \quad x_3^{(2)}, \cdots, \quad x_3^{(10)}.$$

For any series path with random variables

$$x_1^{(i)}, \quad x_2^{(i)}, \quad x_3^{(i)}, \qquad i = 1, 2, \cdots, 10,$$

the linking and blocking probabilities for this path are respectively $q_1^{(i)} q_2^{(i)} q_3^{(i)}$ and $1 - q_1^{(i)} q_2^{(i)} q_3^{(i)}$ where $q_1^{(i)} = Pr[x_1^{(i)} = 1]$ and similarly for $q_2^{(i)}$ and $q_3^{(i)}$. Thus, for G,

$$P = \prod_{i=1}^{10} (1 - q_1^{(i)} q_2^{(i)} q_3^{(i)}), \qquad Q = 1 - P$$

Numerically, suppose $q_1^{(i)} = q_2^{(i)} = q_3^{(i)} = \frac{2}{3}$ for all i, then

$$P = \left[1 - \left(\frac{2}{3}\right)^3\right]^{10} = \left(\frac{19}{27}\right)^{10} = 0.0298\cdots$$

This example illustrates the fact that linking and blocking probabilities of a series-parallel graph can be found in a routine manner. Although the same can be said for nonseries-parallel graphs, the computational difficulties involved are of a different nature.

Let G be a probability linear-graph and let G^* be its canonical form. Denote the link random variables of G^* by $y^{(1)}, y^{(2)}, \cdots$ and let $Y^{(i)}$ be the event $y^{(i)} = 1$. Then clearly, the linking probability Q of G is the probability of the union of events $Y^{(i)}$ or

$$Q = Pr\left[\bigcup_i Y^{(i)}\right] \tag{2.19}$$

This last probability can be found in a standard manner (cf. Reference 8, Chapter 4) since the joint probabilities $Pr[\bigcap_j Y^{(i_j)}]$ are readily obtained from the linking probabilities of the links of G.

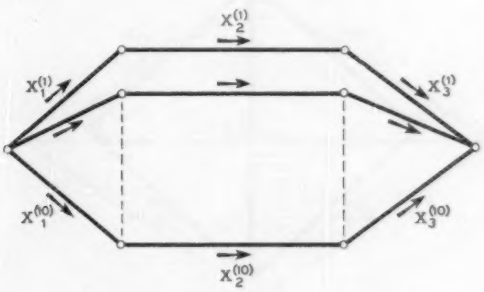


FIG. 2.3 — A series-parallel graph.

A real computational difficulty arises at this point. If the number of directed paths of G (or G^*) is not small, then the expansion of Q in terms of the aforementioned joint probabilities becomes formidable.

Example 2.2. In Figs. 2.4 and 2.5 are a non-series-parallel graph G and its canonical form G^*, respectively. In this case,

$$Q = Pr[Y^{(1)} \cup Y^{(2)} \cup Y^{(3)}] = \sum_{i=1}^{3} Pr[Y^{(i)}] - \sum_{\substack{i,j=1 \\ i<j}}^{3} Pr[Y^{(i)} \cap Y^{(j)}] + \sum_{\substack{i,j,k=1 \\ i<j<k}}^{3} Pr[Y^{(i)} \cap Y^{(j)} \cap Y^{(k)}] \tag{2.20}$$

Since $Pr[Y^{(i)} \cap Y^{(j)} \cap \cdots]$ can be found directly from G (e.g. $Pr[Y^{(1)} \cap Y^{(2)}] = q^{(1)}q^{(2)}q^{(3)}q^{(5)}$), we get

$$\begin{aligned} Q = {} & (q^{(1)}q^{(2)} + q^{(2)}q^{(3)}q^{(5)} + q^{(3)}q^{(4)}) \\ & - (q^{(1)}q^{(2)}q^{(3)}q^{(5)} + q^{(1)}q^{(2)}q^{(3)}q^{(4)} + q^{(2)}q^{(3)}q^{(4)}q^{(5)}) \\ & + (q^{(1)}q^{(2)}q^{(3)}q^{(4)}q^{(5)}) \end{aligned} \tag{2.21}$$

In particular, if $q^{(1)} = q^{(2)} = q^{(3)} = q^{(4)} = q^{(5)} = q$, then

$$Q = 2q^2 + q^3 - 3q^4 + q^5 \tag{2.22}$$

2.5 *Generating Functions for Linking Probabilities.*

Let G be a graph with N links and let the links be designated by $X^{(1)}, X^{(2)}, \cdots, X^{(N)}$. Let B be the set of all directed paths of G. Then each $\beta_j \in B$ is composed of links $X^{(j_1)}, X^{(j_2)}, \cdots, X^{(j_{n_j})}$ in series. Let us agree to denote the directed path β_j by the formal product

$$X^{(j_1)} \cdot X^{(j_2)} \cdot \cdots \cdot X^{(j_{n_j})}.$$

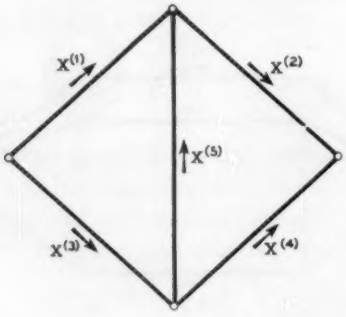


FIG. 2.4 — A nonseries-parallel graph.

In this formal product, the X's are considered as undefined real numbers and are manipulated as such *except* for the *reduction rule*

$$X^{(j)} \cdot X^{(k)} = X^{(j)} \quad \text{if} \quad j = k. \tag{2.23}$$

Otherwise, the operation $\cdot$ is ordinary multiplication.

Definition 2.4. A function $\mathcal{Q}$ is said to be the *generating function* of G if

$$\mathcal{Q}(s) = 1 - \prod_{B}^{*} (1 - X^{(j_1)} \cdot X^{(j_2)} \cdot \cdots \cdot X^{(j_{n_j})} \cdot s) \tag{2.24}$$

where Π^* denotes formal product under $\cdot$ such that *the reduction rule (2.23) is always carried out.*

Example 2.3. Let us go back to Fig. 2.4. In this case, there are three directed paths $X^{(1)} \cdot X^{(2)}$, $X^{(3)} \cdot X^{(5)} \cdot X^{(2)}$ and $X^{(3)} \cdot X^{(4)}$ so that

$$\begin{aligned} \mathcal{Q}(s) = {} & (X^{(1)} \cdot X^{(2)} + X^{(3)} \cdot X^{(4)} + X^{(2)} \cdot X^{(3)} \cdot X^{(5)})s \\ & - (X^{(1)} \cdot X^{(2)} \cdot X^{(3)} \cdot X^{(5)} + X^{(1)} \cdot X^{(2)} \cdot X^{(3)} \cdot X^{(4)} \\ & + X^{(2)} \cdot X^{(3)} \cdot X^{(4)} \cdot X^{(5)})s^2 + (X^{(1)} \cdot X^{(2)} \cdot X^{(3)} \cdot X^{(4)} \cdot X^{(5)})s^3 \end{aligned} \tag{2.25}$$

Note that if, in (2.25), each $X^{(j)}$ is replaced by the real number $q^{(j)}$ and s is set equal to 1, the resulting value of the generating function $\mathcal{Q}$[denoted by $\mathcal{Q}^*(1)$] is

$$\begin{aligned} \mathcal{Q}^*(1) = {} & (q^{(1)}q^{(2)} + q^{(2)}q^{(3)}q^{(5)} + q^{(3)}q^{(4)}) \\ & - (q^{(1)}q^{(2)}q^{(3)}q^{(5)} + q^{(1)}q^{(2)}q^{(3)}q^{(4)} + q^{(2)}q^{(3)}q^{(4)}q^{(5)}) \\ & + (q^{(1)}q^{(2)}q^{(3)}q^{(4)}q^{(5)}) \end{aligned} \tag{2.26}$$

which is precisely the linking probability Q for the graph found previously (Example 2.2). In fact, this relation holds true in general.

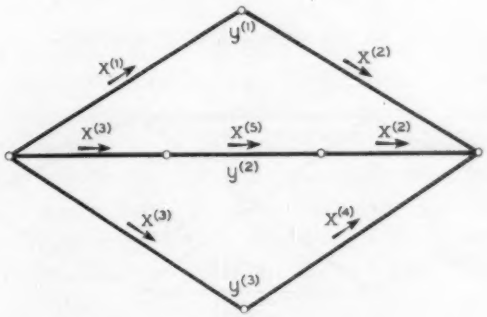


FIG. 2.5 — Canonical form of graph shown in Figure 2.4.

Theorem 2.2. Let G be a probability linear-graph and $\mathcal{Q}$ its generating function. Then $\mathcal{Q}^*(1)$ is the linking probability Q of G.

Proof. Let B be the set of all directed paths of G and let $y^{(i)}$ be the random variables associated with these directed paths. Let $Y^{(i)}$ be the event $y^{(i)} = 1$. Then the coefficient of s^r in $\mathcal{Q}^*(s)$ is

$$\sum_{\substack{i_1, i_2, \ldots, i_r \\ i_1 < i_2 < \ldots < i_r}} Pr(Y^{(i_1)} \cap Y^{(i_2)} \cap \ldots \cap Y^{(i_r)})$$

except for a possible change in sign. Thus

$$\begin{aligned} \mathcal{Q}^*(1) = \sum_i Pr(Y^{(i)}) - \sum_{\substack{i_1, i_2 \\ i_1 < i_2}} Pr(Y^{(i_1)} \cap Y^{(i_2)}) + \ldots \\ \pm \sum_{\substack{i_1, i_2, \ldots, i_n \\ i_1 < i_2 < \ldots < i_n}} Pr(Y^{(i_1)} \cap Y^{(i_2)} \cap \ldots \cap Y^{(i_n)}) = Q \end{aligned} \tag{2.27}$$

This completes the proof.

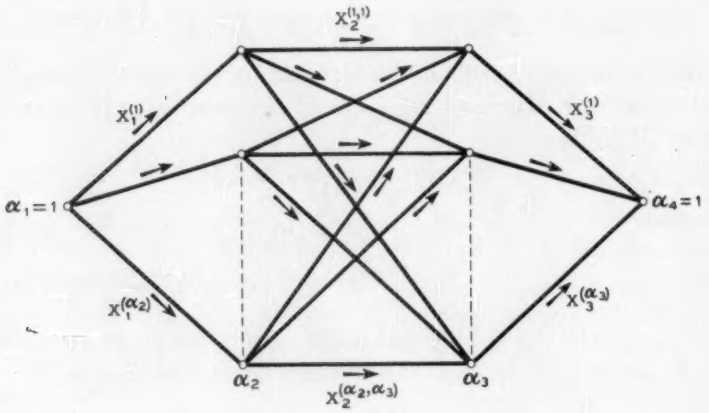


FIG. 2.6 — A general four-stage graph.

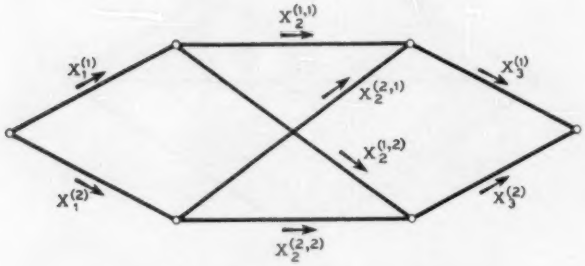


FIG. 2.7 — A simple four-stage graph.

2.6 *Some 4-Stage Graphs*

By a four-stage graph is meant a graph of the form shown in Fig. 2.6 where the graph has four stages of nodes with $\alpha_1 = 1$, α_2, α_3 and $\alpha_4 = 1$ nodes in each stage. It is clear from Fig. 2.6 that there are α_2 left links, $\alpha_2\alpha_3$ middle links and α_3 right links.

The linking probability of a four-stage graph can theoretically be found by a direct application of Theorem 2.2. The computation is simplified, however, if additional information is given. For instance, consider

Example 2.4. Let G be the four-stage graph shown in Fig. 2.7 with $\alpha_2 = \alpha_3 = 2$ where

$$\left.\begin{aligned} q_1^{(1)} &= q_1^{(2)} = q_1 \\ q_2^{(1,1)} &= q_2^{(1,2)} = q_2^{(2,1)} = q_2^{(2,2)} = q_2 \\ q_3^{(1)} &= q_3^{(2)} = q_3 \end{aligned}\right\} \tag{2.28}$$

Then we may write

$$\begin{aligned} \mathcal{Q}(s) = 1 - (1 - X_1^{(1)} \cdot X_3^{(1)} \cdot q_2 s) \cdot (1 - X_1^{(1)} \cdot X_3^{(2)} \cdot q_2 s) \\ \cdot (1 - X_1^{(2)} \cdot X_3^{(1)} \cdot q_2 s) \cdot (1 - X_1^{(2)} \cdot X_3^{(2)} \cdot q_2 s) \end{aligned} \tag{2.29}$$

so that

$$Q = \mathcal{Q}^*(1) = 4q_1q_2q_3 - 2q_1q_2^2q_3(q_1 + q_1q_3 + q_3) + 4q_1^2q_2^3q_3^2 - q_1^2q_2^4q_3^2$$

The same procedure can be applied to all such 4-stage graphs.

Example 2.5. In the special cases where the occupancies of the left, middle and right links are equal respectively to p_1, p_2 and p_3 (Example 2.4 is one of these), the following expression* has been obtained by D. H. Evans for the blocking probability P of a 4-stage graph:

$$P = \sum_{k=0}^{a2} \binom{\alpha_2}{k} (1 - p_1)^k p_1^{(\alpha_2 - k)} [(1 - p_3)p_2^k + p_3]^{\alpha_2}$$

This formula is of considerable value in a study of blocking in six-stage switching networks by the author and D. H. Evans (unpublished).

In general, if the number of nodes of a four-stage graph is not small, it would be unrealistic timewise to expand $\mathcal{Q}(s)$ to yield the linking probability Q of the graph. In such cases no good theoretical method has been found which will conveniently yield the values of Q. It is possible, however, with the aid of Theorem 2.1, to devise an experiment which will yield good approximations to Q without undue labor. In fact, this remark applies to all probability linear-graphs.

* This expression is derived from a counting procedure different from the approach outlined here. Using still another approach, an extension of this result to a more general class of 4-stage networks has been obtained by M. Goldman.

We shall now proceed to describe switching networks as probability linear-graphs.

3. SWITCHING NETWORKS AS PROBABILITY LINEAR-GRAPHS

3.1 *Probability of Blocking.*

By a switching network is meant a crosspoint network in which each input of the network can be connected to any output of the network by the operation of appropriate sequences of crosspoints. A block diagram representation of such networks is shown in Fig. 3.1. The first stage of crosspoint switches has L_0 inputs and L_1 outputs which are inputs to the second stage of switches. Thus these L_1 leads link stages 1 and 2 and are therefore called L_1-links. Similarly, the L_i-links $0 < i < s$ link stages L_i and L_{i+1} and the L_s links are the outputs of the network which has s stages.

Let L be a switching network with inputs and outputs indexed by i and j; $i = 1, 2, \cdots, L_0$, $j = 1, 2, \cdots, L_s$. Denote by $P(i, j)$ the probability of all paths from input i to output j busy. Then the blocking probability P of L is defined here as*

$$P = \frac{1}{L_0 L_s} \sum_{i=1}^{L_0} \sum_{j=1}^{L_s} P(i, j) \tag{3.1}$$

Similarly the linking probability Q of L is

$$Q = 1 - P \tag{3.2}$$

For each pair (i, j), $P(i, j)$ is found as the blocking probability of a corresponding probability linear-graph.

In this section, we assume $P(i, j)$ to be independent of i and j and each crosspoint switch (switch with arrays of crosspoints) in the network to be of nonblocking type.† There is no essential loss of generality in the first condition; although without it, the evaluation of P would become more tedious.

It remains for us to present the correspondence between a switching network L and a probability linear-graph G. This description is best given by examples. Choose an input i and an output j of L. Then the paths from i to j would involve crosspoint switches and links between

* This is one of several possible definitions of blocking probability of a network.

† Here we restrict our attention to networks consisting of switches of nonblocking type (e.g., square switches or switches with more outputs than inputs). Thus, "concentration" switches are excluded. The analysis of networks with concentrating stages involves deeper insight and will not be considered in this paper.

pairs of these switches. Represent the switches by nodes and the links by directed branches. The resulting graph G is the graph corresponding to L.

We consider G to be the model of L from which pertinent characteristics of L can be extracted. Thus, we are identifying such quantities as blocking probability, mean retrial time etc., of L with those of G.

Example 3.1. The 4-stage switching distribution network L shown in Fig. 3.2 is common (with modifications) in telephone central offices. Its corresponding probability linear-graph is shown in Fig. 3.3, with all link occupancies p. The blocking probability P of L is therefore (see Example 2.1).

$$P = [1 - (1 - p)^3]^{10} \tag{3.3}$$

Equation (3.3) is the simplest one of several formulas sometimes known as Kittredge-Molina formulas for crosspoint networks.

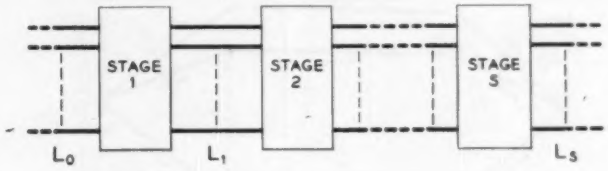


FIG. 3.1 — Block diagram of a switching network.

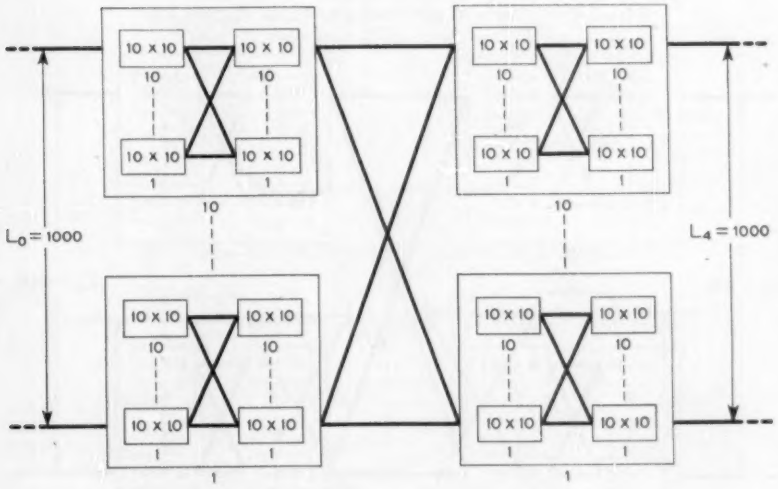


FIG. 3.2 — A four-stage switching network.

Example 3.2. Consider the partially-equipped, four-stage distribution network, first suggested by C. A. Lovell, (Fig. 3.4) which would grow with relative ease into the full distribution network of Example 3.1. In this example, we prefer to overlook the fact that the third stage switches are not of the non-blocking type. The corresponding graph is shown in Fig. 3.5 and the blocking probability is

$$P_1 = [1 - (1 - p_1)^2(1 - p_2)]^{10} \tag{3.4}$$

To calculate the link occupancies, let A be the offered traffic (in erlangs) in Example 3.1 and for the purpose of comparison, let 0.4A be the offered traffic in Example 3.2. Then

$$p = \frac{A(1 - P)}{1000}, \qquad p_1 = \frac{0.4A(1 - P_1)}{400}, \qquad p_2 = \frac{0.4A(1 - P_1)}{1000} \tag{3.5}$$

where in (3.5), the link occupancies depend themselves on the blocking

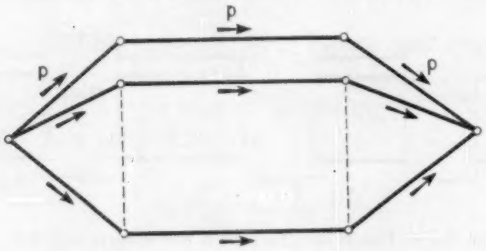


FIG. 3.3 — Graph of network shown in Figure 3.2.

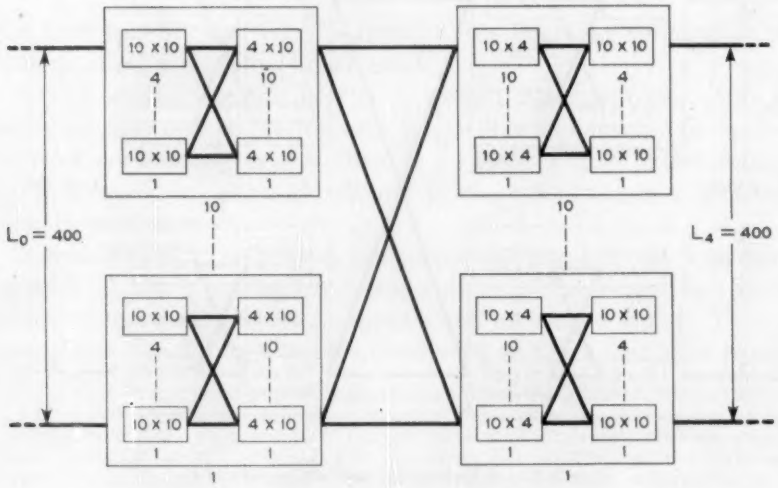


FIG. 3.4 — A partially equipped four-stage network.

probabilities. As a first approximation, the factors $(1 - P)$ and $(1 - P_1)$ may be neglected. A comparison of (3.3) and (3.4) shows then that the blocking of a partially equipped network is less than that of a fully equipped network.

Example 3.3. The switching network shown in Fig. 3.6 is a particular case of a class of non-blocking networks discovered by C. Clos.[10] Its corresponding graph is shown in Fig. 3.7 with blocking probability

$$P = [1 - (1 - p)^2]^{39} \tag{3.6}$$

Since it is known a priori that the actual blocking probability for the network is identically zero, it is interesting to compare P given by (3.6) with zero. We arrive at the following table:

p	P
0.1	7.44×10^{-29}
0.3	3.94×10^{-12}
0.5	1.34×10^{-6}
0.7	0.025
0.9	0.676
0.95	0.907
1.0	1.0

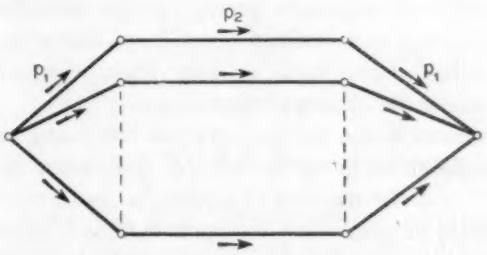


FIG. 3.5 — Graph of network shown in Figure 3.4.

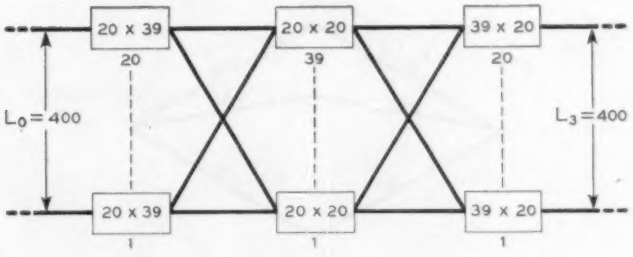


FIG. 3.6 — A three-stage non-blocking network.

This example illustrates the effect of neglecting link dependence; Equation (3.6) gives very high blocking for large p. In practice, however, a value of p greater than 0.5 is rarely encountered so that even in this case the approximation is not unreasonable. It should be remarked that more suitable formulas for computing blocking in three-stage networks have been developed by M. Karnaugh.

3.2 *Applications to Equivalent Crosspoint Minimization.*

As an application of the concepts developed here, we shall consider the problem of equivalent crosspoint minimization for switching networks.

The case in mind is a four-stage network in which we assume:

A-1. The network is symmetrical with respect to its inputs and outputs; (i.e., the network configuration remains fixed when the inputs and outputs of the network are interchanged).

A-2. Switches within each stage of the network are of identical size.

A-3. Stages 1 and 2 are combined into β identical frames (called primary frames) no two of which are interconnected.

A-4. The number of outputs in a switch in the first stage is equal to the number of switches in each frame in the second stage, with each output connected to a distinct switch in that frame.

A-5. The number of outputs in a switch in the second stage is β, with each output connected to a distinct (secondary) frame and with outputs on distinct switches in the same primary frame connected to distinct switches in the same (secondary) frame.

Under these assumptions, the network has the configuration shown in Fig. 3.8. The problem is, given the offered traffic, the maximum allowable blocking P' and the number of inputs L_0, determine x, y, α and β so that the number of equivalent crosspoints C_0 is a minimum where C_0 is given by

$$C_0 = C + k_0 L_0 + 2k_1 L_1 + 2k_2 L_2 \tag{3.7}$$

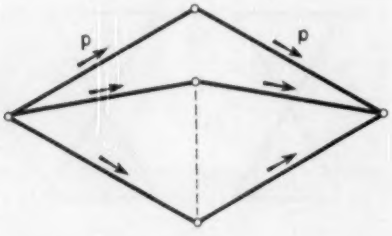


FIG. 3.7 — Graph of the three-stage non-blocking network.

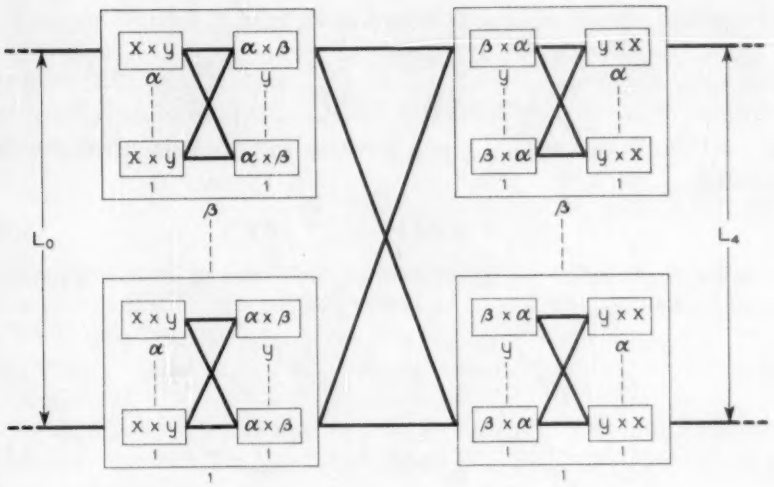


FIG. 3.8 — A general four-stage switching network.

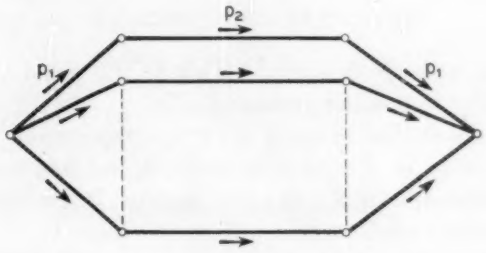


FIG. 3.9 — Graph of the general four-stage switching network.

in which k_0, k_1 and k_2 are given constants which evaluate relative network cost effects and C is the number of crosspoints in the network.

The corresponding graph shown in Fig. 3.9 consists of y parallel paths so that the blocking probability for the network is

$$P = (1 - q_1^2 q_2)^y \tag{3.8}$$

As a first approximation, we have

$$q_1 = 1 - \frac{A'}{\alpha\beta y}, \qquad q_2 = 1 - \frac{A'}{\beta^2 y} \tag{3.9}$$

where

$$A' = A(1 - P') \tag{3.10}$$

and A is the offered erlangs. It then follows that

$$C_0 = L_0(k_0 + 2y) + \frac{2(A')^{3/2}}{(1 - q_1)(1 - q_2)^{1/2}y^{1/2}} + \frac{2k_1A'}{1 - q_1} + \frac{2k_2A'}{1 - q_2} \quad (3.11)$$

Our problem is to find q_1, q_2 and y which will minimize C_0 under the condition

$$(1 - q_1^2q_2) - (P')^{1/y} = 0 \quad (3.12)$$

Using the Lagrangian multiplier method, we arrive at the following system of four equations

$$\frac{(A')^{3/2}}{(1 - q_1)^2(1 - q_2)^{1/2}y^{1/2}} + \frac{k_1A'}{(1 - q_1)^2} = \lambda q_1q_2 \quad (3.13)$$

$$\frac{(A')^{3/2}}{(1 - q_1)(1 - q_2)^{3/2}y^{1/2}} + \frac{k_2A'}{(1 - q_2)^2} = \lambda q_1^2 \quad (3.14)$$

$$\frac{(A')^{3/2}}{(1 - q_1)(1 - q_2)^{1/2}y^{3/2}} - \lambda\frac{(P')^{1/y}\ell nP'}{y^2} = 2L_0 \quad (3.15)$$

$$1 - q_1^2q_2 - (P')^{1/y} = 0 \quad (3.16)$$

in which the unknowns to be solved for are q_1, q_2, y and λ. These can be found numerically in specific problems.

It should be noted that in using the graph representation to find the blocking probability of a switching network, we had tacitly assumed that the switches should be of non-blocking type. But solutions to (3.13)–(3.16) do not guarantee that this will be the case. Thus, the values of x, y, α and β obtained from this minimization process should be regarded as approximations.

We now consider the following special case which is of independent interest. In equation (3.7), let $k_0 = k_1 = k_2 = 0$. This situation corresponds physically to the interpretation that the link costs are negligible as compared to the crosspoint costs. Equations (3.13)–(3.16) then become

$$\frac{(A')^{3/2}}{(1 - q_1)^2(1 - q_2)^{1/2}y^{1/2}} = \lambda q_1q_2 \quad (3.17)$$

$$\frac{(A')^{3/2}}{(1 - q_1)(1 - q_2)^{3/2}y^{1/2}} = \lambda q_1^2 \quad (3.18)$$

$$2L_0 - \frac{(A')^{3/2}}{(1 - q_1)(1 - q_2)^{1/2}y^{3/2}} + \lambda\frac{(P')^{1/y}\ell nP'}{y^2} = 0 \quad (3.19)$$

$$1 - q_1^2q_2 - (P')^{1/y} = 0. \quad (3.20)$$

From (3.17) and (3.18) it turns out that

$$q_1 = q_2 = q \tag{3.21}$$

so that the problem of solving four equations is now reduced to that of solving a single equation

$$\frac{(A')^{3/2}}{(1-q)^{3/2}y^{3/2}}\left[1 - \frac{(1-q^3)\,\ell n\,(1-q^3)}{q^2(1-q)}\right] = 2L_0. \tag{3.22}$$

It is interesting to note that in this case, as an immediate consequence of the minimization process, the second and third stage switches turn out to be square switches and hence are of non-blocking type. We remark also that a similar minimization process may be applied to three-stage networks.

Example 3.4. Suppose it is desired to design a minimum crosspoint, four-stage network with $A = 400$ erlangs, $L_0 = 800$ input lines and $P' = 0.01$. Solving Equation (3.22), we find

$$y = 15 \qquad q = 0.642$$

Therefore, from Equations (3.9) and (3.10), we get

$$\alpha = \beta = 8.65 \cong 9, \qquad x \cong 10.$$

For these values of α, β, x and y, we find

$$C = 46{,}200 \text{ crosspoints.}$$

4. RETRIAL AND CONNECTION-TIME DISTRIBUTIONS

4.1 *Retrial Distribution.*

Given a switching network, it is of interest to know whether a path which is blocked at some given time can be established some time later and with what degree of success. More precisely, let L be a switching network and G its corresponding graph. By the *retrial distribution* (denoted by $P_{00}(t)$) of G is meant the conditional probability*

$$P_{00}(t) = Pr[\varphi_{\tau+t} = 0 \mid \varphi_\tau = 0]$$

where φ_t is the connection function of G.

In order to find $P_{00}(t)$, we consider the following process† for the link random variables:

1. The link random variables are mutually independent with common probability distributions.

* The notation $P_{\delta_1\delta_2}(t)$ is used for conditional probabilities whereas the notation $P_t(\delta_1, \delta_2)$ is used for joint probabilities.

† This is a special case of stochastic processes discussed in Reference 1.

2. If X is a link random variable, the conditional probability of a change from $x = 1$ to $x = 0$ during $(t, t + h)$ is $\lambda h + 0(h)$, from $x = 0$ to $x = 1$ during $(t, t + h)$ is $\mu h + 0(h)$. The conditional probability of more than one change in $(t, t + h)$ is $0(h)$; (i.e. the link holding and idle time lengths are exponentially distributed with parameters μ and λ; cf. Reference 8, Chapter 17).

We then obtain the following system of four differential equations for each link random variable.

$$\begin{aligned} P_{00}'(t) &= -\mu P_{00}(t) + \lambda P_{01}(t)\cdot \\ P_{01}'(t) &= -\lambda P_{01}(t) + \mu P_{00}(t)\cdot \\ P_{10}'(t) &= -\mu P_{10}(t) + \lambda P_{11}(t)\cdot \\ P_{11}'(t) &= -\lambda P_{11}(t) + \mu P_{10}(t)\cdot \end{aligned} \tag{4.1}$$

with initial conditions

$$P_{00}(0) = P_{11}(0) = 1, \qquad P_{10}(0) = P_{01}(0) = 0$$

The solution to this system of equations is (cf. Reference 1)

$$\left.\begin{aligned} P_{00}(t) &= \frac{\lambda + \mu e^{-(\lambda+\mu)t}}{\lambda + \mu} = p + qe^{-(\lambda+\mu)t} \\ P_{01}(t) &= \frac{\mu - \mu e^{-(\lambda+\mu)t}}{\lambda + \mu} = q - qe^{-(\lambda+\mu)t} \\ P_{10}(t) &= \frac{\lambda - \lambda e^{-(\lambda+\mu)t}}{\lambda + \mu} = p - pe^{-(\lambda+\mu)t} \\ P_{11}(t) &= \frac{\mu + \lambda e^{-(\lambda+\mu)t}}{\lambda + \mu} = q + pe^{-(\lambda+\mu)t} \end{aligned}\right\} \tag{4.2}$$

where p is the occupancy of the link and $q = 1 - p$. Since

$$\left.\begin{aligned} P_t(1, 1) &= 1 - 2P(0) + P_t(0, 0) \\ P_t(0, 0) &= 2P(0) + P_t(1, 1) - 1 \\ P_t(1, 0) &= P_t(0, 1) = P(0) - P_t(0, 0) = P(1) - P_t(1, 1) \end{aligned}\right\} \tag{4.3}$$

retrial distributions for series-parallel graphs can be found directly from (4.3) or (2.7)–(2.10) and (2.11)–(2.14).

Example 4.1. Consider the switching network shown in Fig. 3.2 the corresponding graph of which was shown in Fig. 3.3. For this graph we find, for each path, [say path (i)]

$$P_t^{(i)}(1, 1) = P^{(i)}(1)P_{11}^{(i)}(t) = q^3(q + pe^{-(\lambda+\mu)t})^3 \tag{4.4}$$

Hence, for the entire graph, from (4.3) and (3.3)

$$P_t(0, 0) = [2(1 - q^3) + q^2(q + pe^{-(\lambda+\mu)t})^3 - 1]^{10} \tag{4.5}$$
$$= (1 - 2q^3 + [q(q + pe^{-(\lambda+\mu)t})]^3)^{10}$$

To fix ideas, let the "average holding time" $1/\mu = 200$ sec., and link occupancy $p = \frac{1}{3}$. Then, since

$$\lambda q = \mu p, \tag{4.6}$$

we get

$$\lambda = \frac{1}{400}, \qquad q = \frac{2}{3} \tag{4.7}$$

For these values of μ and p, we get

$$P_t(0, 0) = \left[\frac{11}{27} + \frac{8}{27}\left(\frac{2}{3} + \frac{1}{3}e^{-(3/400)t}\right)^3\right]^{10} \tag{4.8}$$

4.2. *Mean Retrial Time*

Let us now consider the following problem. We begin with the hypothesis that all paths from a given input to a given output of a switching network are busy initially. If a retrial is made every x seconds until one path is found free, what is the expected number of seconds $\bar{t}_x$ for the establishment of a path?

To solve this problem, let G be the graph corresponding to the switching network with connection function φ_t with probability distributions P, Q, $P_t(0, 0)$, $P_t(0, 1)$, $P_t(1, 0)$ and $P_t(1, 1)$. We use the notation $P_{\delta_1\delta_2}(t)$ to mean the transition probability of $\varphi_t = \delta_2$ at time t given $\varphi_t = \delta_1$ initially. Let $m_x(n)$ be the probability of success at the n^{th} trial, given that all paths are busy initially. Then

$$\bar{t}_x = \sum_{k=1}^{\infty} kxm_x(k) \tag{4.9}$$

Since

$$m_x(k) = (P_{00}(x))^{k-1}P_{01}(x); \qquad k = 1, 2, \cdots \tag{4.10}$$

we find

$$\bar{t}_x = \sum_{k=1}^{\infty} kx(P_{00}(x))^{k-1}P_{01}(x) = \frac{x}{1 - P_{00}(x)} = \frac{xP}{P - P_x(0, 0)} \tag{4.11}$$

where P is the blocking probability of the network.

We shall call $\bar{t}_x$ the *mean retrial time* (in interval of x seconds). In

particular, if retrials are made in arbitrarily small time interval, the limit $\lim_{x\to 0}\bar{t}_x$, when existent, is called the *limiting retrial time* $\bar{t}$. Physically, $\bar{t}$ is the expected number of seconds the first path becomes free given that all paths are busy initially. From (4.11), we find

$$\bar{t} = \lim_{x\to 0} \frac{xP}{P - P_x(0, 0)} = -\frac{P}{\left[\frac{d}{dx} P_x(0, 0)\right]_{x=0}} \tag{4.12}$$

Example 4.2. Going back to Example 4.1, $P_x(0, 0)$, which is the joint probability of the network blocked initially and also blocked x seconds later, was given by (4.8):

$$P_x(0, 0) = \left[\frac{11}{27} + \frac{8}{27}\left(\frac{2}{3} + \frac{1}{3} e^{-(3/400)x}\right)^3\right]^{10}$$

Differentiating $P_x(0, 0)$ with respect to x and setting $x = 0$, we find the limiting retrial time as given by (4.12) is 31.67 seconds. On the other hand, if a retrial is made once every second (that is, $x = 1$), the mean retrial time as given by (4.11) is 32.26 seconds.

We remark that together with the method of Section 2, the same procedure can be applied to non-series-parallel graphs such as those shown in Figures 2.4 and 2.7.

4.3 *Connection-Time Distributions*

Up to now we have not paid attention to the physical structure of the crosspoints in a switching network. If the crosspoints are made up of active elements such as gas diodes, the following problem then presents itself. It is known that it takes a certain time for a gas tube to break down after a voltage higher than the breakdown voltage is applied. In a switching network, the establishment of a path in general involves the breakdown of several such crosspoints in series and it is important that the breakdown time of a path must be reasonably short lest the system bogs down from this inherent delay.

There is experimental evidence that the breakdown time of a gas tube has roughly an exponential distribution. Thus, if X is a random variable representing the breakdown time of a tube (for a fixed applied voltage), then we postulate

$$Pr[X \leqq x] = 1 - e^{-\alpha x} \tag{4.13}$$

where

$$E(X) = \frac{1}{\alpha} \tag{4.14}$$

We shall assume that the tubes behave independently and have identical breakdown time characteristics and that the voltage across each tube before breakdown remains a fixed constant regardless of the behavior of the other tubes. The reason for imposing the latter condition is clear since the mean breakdown time $1/\alpha$ depends upon the voltage applied.

Let L be a gas tube switching network and G its corresponding graph with connection function φ_t. Let ψ be a random variable representing the breakdown time of G. We define the *connection-time distribution* Φ of L to mean the joint probability.

$$\Phi = Pr[\varphi_t = 1 \text{ and } \psi \leqq t] \tag{4.15}$$

i.e., the joint probability of finding at least one idle path and the breakdown time of some idle path less than or equal to t. It is also desirable to denote by Ψ the probability

$$\Psi = 1 - \Phi = Pr[\varphi_t = 0 \text{ or } (\varphi_t = 1 \text{ and } \psi \geqq t)] \tag{4.16}$$

In a graph G, the tubes are represented by nodes of G. The number of nodes in a path in G then represents the number of tubes in series in the path.

To be concrete, let us now consider the network shown in Fig. 3.2 with its associated graph shown in Fig. 3.3. We shall study two methods of establishing paths through the network.

1. *End-Matching.* A voltage is applied to both ends of the network simultaneously (across a single input and a single output). We wish to find Φ (or Ψ) of the network.

First, let us digress to state a useful lemma the proof of which can be found in probability texts.

Lemma 4.1. Let $X_1, X_2, \cdots, X_n$ be mutually independent random variables with common distribution

$$Pr[X_i \leqq x] = 1 - e^{-\alpha x}; \qquad x \geqq 0 \tag{4.17}$$

then the distribution of the sum $X_1 + X_2 + \cdots + X_n$ is

$$Pr(X_1 + X_2 + \cdots + X_n \leqq t) = 1 - e^{-\alpha t} \sum_{j=0}^{n-1} \frac{(\alpha t)^j}{j!} \tag{4.18}$$

Reproducing the graph, Fig. 3.3, we have, for a single path

Pr[(path busy) or (path idle and breakdown time $\geqq t$)]

$$= Pr[\text{path busy}] + Pr[\text{path idle and breakdown time} \geqq t] \tag{4.19}$$

Now a path is idle only if all three links making up the path are idle so

that, from Lemma 4.1,*

$$\Pr[\text{path idle and breakdown time} \geqq t] = Pr[\text{path idle}] \cdot Pr_{\text{given path idle}}[\text{breakdown time} \geqq t] = q^3 e^{-\alpha t} \sum_{j=0}^{3} \frac{(\alpha t)^j}{j!} \tag{4.20}$$

Hence, (4.19) is given by the expression

$$(1 - q^3) + q^3 e^{-\alpha t} \sum_{j=0}^{3} \frac{(\alpha t)^j}{j!} \tag{4.21}$$

For ten independent paths in parallel, it is clear

$$\begin{aligned} \Psi &= Pr[\varphi_t = 0 \text{ or } (\varphi_t = 1 \text{ and } \psi \geqq t)] \\ &= \{Pr[(\text{one path busy}) \text{ or } (\text{one path idle and breakdown time} \geqq t)]\}^{10} \\ &= \left[(1 - q^3) + q^3 e^{-\alpha t} \sum_{j=0}^{3} \frac{(\alpha t)^j}{j!}\right]^{10} \end{aligned} \tag{4.22}$$

and

$$\Phi = 1 - \Psi = 1 - \left[(1 - q^3) + q^3 e^{-\alpha t} \sum_{j=0}^{3} \frac{(\alpha t)^j}{j!}\right]^{10} \tag{4.23}$$

We shall defer numerical examples in order to arrive at a comparison with the next case.

2. *Center-Matching.* In this method, the marks propagate in both directions and are matched in the center links (see Fig. 4.1).

In this case, we need to consider the left side and right side breakdown times individually and then combine them. For a single path, we may write

$$Pr[\text{path idle and breakdown time} \leqq t] = \text{I} \cdot \text{II} \tag{4.24}$$

where

$$\left.\begin{aligned} \text{I} &= Pr[\text{path idle}] \\ \text{II} &= Pr_{\text{given path idle}}[\text{breakdown time} \leqq t] \end{aligned}\right\} \tag{4.25}$$

The conditional probability II can be written as a product

$$\begin{aligned} \text{II} = \{&Pr_{\text{given path idle}}[\text{breakdown time of left side} \leqq t]\} \\ &\cdot \{Pr_{\text{given path idle}}[\text{breakdown time of right side} \leqq t]\}, \end{aligned} \tag{4.26}$$

* We shall use the notations $Pr[B \mid A]$ and $Pr_{\text{given event } A}$ [event B] interchangeably.

so that, from Lemma 1,

$$\Pi = \left(1 - e^{-\alpha t} \sum_{j=0}^{1} \frac{(\alpha t)^j}{j!}\right)^2 \tag{4.27}$$

Combining as before, we obtain

$$\Psi = \left[1 - q^3 \left(1 - e^{-\alpha t} \sum_{j=0}^{1} \frac{(\alpha t)^j}{j!}\right)^2\right]^{10} \tag{4.28}$$

$$\Phi = 1 = \Psi \tag{4.29}$$

We shall consider the following numerical example. For a 10,000-line telephone central office, it is reasonable to assume, from the point of view of control circuits, that the connection time of a call should not exceed 5 milliseconds. Using this value for t and letting $q = 0.7$, we find for several values of mean tube breakdown time (0.5, 1, 5 and 10 milliseconds) the probabilities Ψ shown in Table 4.1.

This table shows that for the network in question, in order to meet the 5 millisecond time requirement, the tubes must have a mean breakdown time of much less than 1 millisecond to insure reasonable blocking.

4.4 *Remarks and Conclusions.*

From the discussion of the last section it is clear that, for a gas tube switching network, blocking probability alone is insufficient as a design

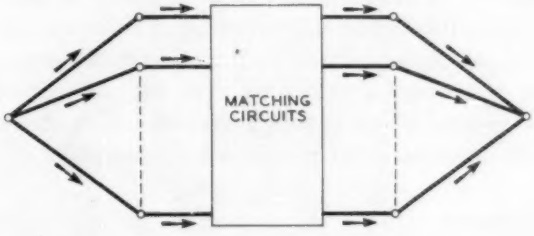


FIG. 4.1 — Graph of four-stage gas tube network with center matching.

TABLE 4.1

Mean Tube Breakdown Time in Milliseconds $1/\alpha$	αt	End-Matching Ψ	Center-Matching Ψ	Blocking Probability P
0.5	10	0.0158	0.0153	0.015
1	5	0.0548	0.0225	0.015
5	1	0.937	0.785	0.015
10	0.5	0.99	0.97	0.015

criterion, since it neglects completely the control circuit requirements. It is more natural to use the joint probability Ψ (which is the probability of either no path available or, when there are paths available, the breakdown time exceeds t) as the generalized blocking for design considerations. Table 4.1 shows how radically different Ψ can be from the blocking probability P.

We may consider the converse situation. Suppose, for the 10,000-line office considered before, in addition to restricting the connection time to not over 5 milliseconds, we further demand that the generalized blocking Ψ must not be higher than some fixed value (say 0.02). Then, from the discussion given here, it is possible to determine the maximum gas tube breakdown time allowable. Thus, generalized blocking can be used as a criterion in determining the choice of the type of gas tubes suitable for the switching network in question.

The fact that it takes time to establish a path in a gas tube network has a definite bearing on the retrial distribution discussed in Section 4.1. Because of it, the mean retrial time on blocked calls will be modified; that is, the retrial distribution of a network depends in part on the characteristics of the gas tubes used in the network.

In this paper we have attempted to study switching networks in terms of a simplified model. Because of the elementary character of the model chosen, many problems are left unsettled. For example, from the point of view of applicability, one would want to know how our results would alter if the assumption of independence among links is dropped. To go a step further, except for Section 4.3, switching networks in this paper have been considered as isolated entities by themselves. A more realistic study should consist of viewing a switching network together with its associated control circuits which entails great difficulties at this time. Any progress in this direction is, of course, highly desirable.

5. ACKNOWLEDGMENT

The writer wishes to express his indebtedness to E. N. Gilbert and W. S. Hayward, Jr., for general orientation and several helpful discussions on the subject. Thanks are also due to S. H. Washburn and H. N. Seckler for suggestions and criticisms and to M. Goldman and D. H. Evans for many stimulating conversations and arguments.

REFERENCES

1. Brockmeyer, E., Halstrom, H. L. and Jensen, A., The Life and Works of A. K. Erlang, The Copenhagen Telephone Company, Copenhagen, Denmark, 1948.
2. Fry, T. C., Probability and Its Engineering Uses, D. Van Nostrand Company, New York, 1928.

3. Molina, E. C., Application of the Theory of Probability to Telephone Trunking Problems, B.S.T.J. **6**, pp. 461–494, 1927.
4. Jacabaeus, C., Blocking Computations in Link Systems, Ericsson Review, No. 3, pp. 86–100, 1947.
5. Lundkvist, K., Method of Computating the Grade of Service in a Selective State Composed of Primary and Secondary Switches, Ericsson Review, No. 1, pp. 11–17, 1948.
6. Jacobaeus, C., A Study of Congestion in Link Systems, Ericsson Technics, No. 48, 1950.
7. Jensen, A., A Basis for the Calculation of Congestion in Crossbar Systems, Teleteknik, No. 3, 1952.
8. Feller, W., Introduction to Probability Theory and Its Applications, John Wiley & Sons, Inc., New York, 1950.
9. Doob, J. L., Stochastic Processes, John Wiley & Sons, Inc., New York, 1953.
10. Clos, C., A Study of Non-Blocking Switching Networks, B.S.T.J. **32**, pp. 406–424, 1953.

Bell System Technical Papers Not Published in This Journal

ARNOLD, W. O.,[1] and HOEFLE, R. R.[1]
A System Plan for Air Traffic Control Embodying the Cussor-Coordinated Display, Proc. I.R.E., P.G.A.N.E. 2, pp. 14–22, June, 1955.

BECKER, J. A.,[1] and BRANDES, R. G.[1]
On The Adsorption of Oxygen on Tungsten as Revealed in The Field Emission Electron Microscope, J. Chem. Phys., **23,** pp. 1323–1330, July, 1955.

BIONDI, F. J.[1]
The Chemist's Role in Electronics, Research and Engineering, **1,** pp. 16–21, July–Aug., 1955.

BOGERT, B. P.[1]
Some Gyrator and Impedance Inverter Circuits, Proc. I.R.E., **43,** pp. 793–796, July, 1955.

BOORSE, H. A., see Smith, B.

BRANDES, R. G., see Becker, J. A.

BRATTAIN, W. H., see Garrett, C. G. B.

COGSWELL, J. W.[2]
Telephone Employes and Public Relations, Telephony, **149,** p. 15, July 30, 1955.

DARROW, K. K.[1]
Some Current Work at Bell Telephone Laboratories, Physics Today, **8,** pp. 6–13, July, 1955.

---

[1] Bell Telephone Laboratories, Inc.
[2] American Telephone and Telegraph Company.

DIMOND, T. L.[1]
Long Distance Dialing and Automatic Accounting, Proc. Railway Systems and Procedures Assoc., Spring Meeting, pp. 173–183, June, 1955.

FINE, M. E.[1]
Elastic Constants of Germanium between 1.7°K and 80°K, J. Appl. Phys., **26,** pp. 862–863, July, 1955.

FISHER, J. R.,[1] and POTTER, J. F.[1]
Factors Affecting Physical Structure of Dry Pressed Steatite, Am. Cer. Soc., Bull., **34,** pp. 177–181, June, 1955.

FORREST, M. E., JR., see Smith, C. W.

GARRETT, C. G. B.,[1] and BRATTAIN, W. H.[1]
Physical Theory of Semiconductor Surfaces, Phys. Rev., **99,** pp. 376–388, July 15, 1955.

GEBALLE, T. H., see Morin, F. J.

GODDARD, C. T.[1]
Measurement of Surface Flatness of Cathodes for Close Spaced Electron Tubes, Ceramic Age, **65,** pp. 20–21, Apr., 1955.

GULDNER, W. G., see Wooten, L. A.

HAGSTRUM, H. D.[1]
Reinterpretation of Electron Impact Experiments in CO, N_2, NO, and O_2, J. Chem. Phys., **23,** pp. 1178–1179, June, 1955.

HARMS, G. J.[2]
Color-Video Transmission over Intercity Television Networks, Elec. Engg., **74,** pp. 667–670, Aug., 1955.

HAUS, H. A.,[6] and ROBINSON, F. N. H.[1]
The Minimum Noise Figure of Microwave Beam Amplifiers, Proc. I.R.E., **43,** pp. 981–991, Aug., 1955.

---

[1] Bell Telephone Laboratories, Inc.
[2] American Telephone and Telegraph Company.
[6] M.I.T., Cambridge, Mass.

HAYNES, J. R.[1]
New Radiation Resulting from Recombination of Holes and Electrons in Germanium, Letter to the Editor, Phys. Rev., **98,** pp. 1866–1868, June 15, 1955.

HEIDENREICH, R. D.[1]
Thermionic Emission Microscopy of Metals. II. Transformations in Plain Carbon Steels, J. Appl. Phys., **26,** pp. 879–889, July, 1955.

HEIDENREICH, R. D.,[1] and STORKS, K. H.[1]
Note on Electron Diffraction Patterns of CuO, Letter to the Editor, J. Appl. Phys., **26,** p. 1056, Aug., 1955.

HOEFLE, R. R., see Arnold, W. O.

JONES, R. V.[2]
Protection Against Electric Shock From Central Office Communications Equipment, Elec. Engg., **74,** pp. 810–813, Sept., 1955.

KRAMER, H. P.[1]
Note on the Emission of Noise by Supersonic Jets, Letter to the Editor, J. Acous. Soc., **27,** pp. 789–790, July, 1955.

KIRCHER, R. J.[1]
Properties of Junction Transistors, Trans. I.R.E., AU-3, pp. 107–124, July–Aug., 1955.

KOLISS, P. P.[1]
Mechanical Splice Closures for Telephone Cables, Telephony, **149,** pp. 23–24, Aug. 13, 1955.

LAW, J. T.[1]
Adsorption of Gases on a Germanium Surface, J. Phys. Chem., **59,** pp. 543–549, June, 1955.

LINVILL, J. G.[1]
Nonsaturating Pulse Circuits Using Two Junction Transistors, Proc. I.R.E., **43,** pp. 826–834, July, 1955.

---

[1] Bell Telephone Laboratories, Inc.
[2] American Telephone and Telegraph Company.

LUKE, C. L.[1]
Determination of Traces of Boron in Silicon, Anal. Chem., **27,** pp. 1150–1153, July, 1955.

MASON, W. P.[1]
Relaxations in the Attenuation of Single Crystal Lead at Low Temperatures and Their Relation to Dislocation Theory, J. Acous. Soc., **27,** pp. 643–653, July, 1955.

MCAFEE, K. B., JR.[1]
Pulse Technique for Measurement of the Probability of Formation and Mobility of Negative Ions, J. Chem. Phys., **23,** pp. 1435, Aug., 1955.

MEYER, F. T.[1]
Improved Detached-Contact Circuit Drawing, Elec. Engg., **74,** p. 645, Aug., 1955.

MOORE, G. E.,[1] WOOTEN, L. A.,[1] and MORRISON, J.[1]
Excess Ba Content of Practical Oxide Coated Cathodes and Thermionic Emission, J. Appl. Phys., **26,** pp. 943–948, Aug., 1955.

MOORE, G. E., see Wooten, L. A.

MORIN, F. J.,[1] and GEBALLE, T. H.[1]
Electrical Conductivity and Seebeck Effect in $Ni_{0.80}Fe_{2.20}O_4$. , Phys. Rev., **99,** pp. 467–468, July 15, 1955.

MORRISON, J., see MOORE, G. E.

PEARSON, G. L.,[1] and WILLIAM, PAUL[4]
Pressure Dependence of the Resistivity of Silicon, Phys. Rev., **98,** pp. 1755–1756, June 15, 1955.

PEDERSON, D. O.[1]
The Regeneration Analysis of Junction Transistor Multivibrators, Trans. I.R.E., P.G.C.T., CT2, pp. 171–178, June, 1955.

POTTER, J. F., see Fisher, J. R.

---

[1] Bell Telephone Laboratories, Inc.
[4] Harvard University, Cambridge, Mass.

ROBINSON, F. N. H., see Haus, H. A.

SHOCKLEY, W.[1]
Semiconductors (In French), Vide, **10,** pp. 9–26, Mar.–Apr., 1955.

SMITH, B.,[1] and BOORSE, H. A.[5]
Helium II Film Transport. III. The Role of Film Height, Phys. Rev., **99,** pp. 358–367, July 15, 1955.

SMITH, B.,[1] and BOORSE, H. A.[5]
Helium II Film Transport. IV. The Role of Temperature, Phys. Rev., **99,** pp. 368–371, July 15, 1955.

SMITH, C. W.,[2] and FORREST, M. E., JR.[3]
A Fully Selective Telemetering System Employing Telegraph Facilities, A.I.E.E. Commun. and Electronics, **19,** pp. 373–377, July, 1955.

STORKS, K. H., see Heidenreich, R. D.

TREUTING, R. G.[1]
Some Aspects of Slip in Germanium, J. of Metals, Sect. 2, **7,** pp. 1027–1031, Sept., 1955.

VOGEL, F. L., JR.[1]
Dislocations in Low-Angle Boundaries in Germanium, Acta Met., **3,** pp. 245–248, May, 1955.

WALKER, L. R.[1]
Generalizations of Brillouin Flow, Letter to the Editor, J. Appl. Phys., **26,** pp. 780–781, June, 1955.

WALKER, L. R.[1]
Power Flow in Electron Beams, J. Appl. Phys., **26,** pp. 1031–1033, Aug., 1955.

WALKER, L. R.,[1] and WOLONTIS, V. M.[1]
Large Signal Theory of Traveling-Wave Amplifiers, Proc. I.R.E., **43,** pp. 260–277, Mar., 1955.

---

[1] Bell Telephone Laboratories, Inc.
[2] American Telephone and Telegraph Company.
[3] Southern Bell Telephone and Telegraph Company.
[5] Columbia University, New York City.

WARNER, A. W.[1]
Frequency Aging of High-Frequency Plated Crystal Units, Proc. I.R.E., **43,** pp. 790–792, July, 1955.

WEIBEL, E. S.[1]
On Webster's Horn Equation, J. Acous. Soc., **27,** pp. 726–727, July, 1955.

WEISS, M. T.[1]
The Behavior of Ferroxdure at Microwave Frequencies, I.R.E. Convention Record, Part 8, pp. 95–99, July, 1955.

WILLIAM, PAUL, see Pearson, G. L.

WINTRINGHAM, W. T.[1]
Review of NBS Circular 526 Optical Image Evaluation, Physics Today, **8,** p. 14, July, 1955.

WOLONTIS, V. M., see Walker, L. R.

WOOTEN, L. A.,[1] MOORE, G. E.,[1] and GULDNER, W. G.[1]
Measurement of Excess Ba in Practical Oxide Coated Cathodes, J. Appl. Phys., **26,** pp. 937–942, Aug., 1955.

WOOTEN, L. A., see Moore, G. E.

[1] Bell Telephone Laboratories, Inc.

Recent Monographs of Bell System Technical Papers Not Published in This Journal*

ANDERSON, P. W., see Weiss, M. T.

BARSTOW, J. M.
Intercity B-W and Color Television Transmission, Monograph 2466.

BEST, F. S., see Harrower, G. A.

BOGERT, B. P.
Stereophonic Sound Reproduction Enhancement Utilizing the Haas Effect, Monograph 2439.

BOGERT, B. P.
Some Gyrator and Impedance Inverter Circuits, Monograph 2444.

BROWN, S. C., see Rose, D. J.

BUEHLER, E., see Tanenbaum, M.

CETLIN, Miss B. B., see Geller, S.

DAVEY, J. R., HANLEY, F. H., and PURVIS, M. R.
A New Telegraph Serviceboard Using Electronic Circuits, Monograph 2366.

DODGE, H. F.
Interpretation of Engineering Data, Monograph 2375.

EDER, Miss M., WARNER, R., and KEENE, F.
Analysis of a Factorial Experiment on Transistors, Monograph 2438.

FELCH, E. P., AND ISRAEL, J. O.
A Simple Circuit for Frequency Standards Employing Overtone Crystals, Monograph 2401.

* Copies of these monographs may be obtained on request to the Publication Department, Bell Telephone Laboratories, Inc., 463 West Street, New York 14, N. Y. The numbers of the monographs should be given in all requests.

FELKER, J. H.
Performance of TRADIC Transistor Digital Computer, Monograph 2395.

FISHER, J. R., and POTTER, J. F.
Factors Affecting Physical Structure of Dry Pressed Steatite, Monograph 2440.

FRANKE, H. C.
Noise Measurement on Telephone Circuits, Monograph 2410.

GEBALLE, T. H., and HULL, G. W.
Seebeck Effect in Silicon, Monograph 2473.

GELLER, S.
The Crystal Structures of RhTe and $RhTe_2$, Monograph 2405.

GELLER, S.
The Crystal Structure of Co_2Si, Monograph 2441.

GELLER, S., and CETLIN, MISS B. B.
The Crystal Structure of $RhSe_2$, Monograph 2476.

GELLER, S., and SCHAWLOW, A. L.
Crystal Structure and Quadrupole Coupling of BrCn, Monograph 2418.

HANLEY, F. H., see Davey, J. R.

HANNA, O. A.
Laminated Crossarms, Monograph 2443.

HANNAY, N. B., see Tanenbaum, M.

HARROWER, G. A., BEST, F. S., and MACHALETT, A. A.
Shielded Electrical Connection Into Vacuum, Monograph 2417.

HOCHGRAF, L., and WATLING, R. G.
Telephone Lines for Rural Subscriber Service, Monograph 2457.

HOOTON, J. A., and MERZ, W. J.
Ferroelectric Domains in $BaTiO_3$ Crystals, Monograph 2416.

HULL, G. W., see Geballe, T. H.

ISRAEL, J. O., see Felch, E. P.

JAYCOX, E. K.
Spectrochemical Procedure of General Applicability, Monograph 2378.

KEENE, F., see Eder, Miss M.

LAW, J. T.
The Absorption of Gases on a Germanium Surface, Monograph 2437.

MACHALETT, A. A., see Harrower, G. A.

MASON, W. P.
Metal Dislocation Relaxations and the Limiting Shearing Stress, Monograph 2433.

MCSKIMIN, H. J.
Elastic Constants of Single Crystal Cobalt, Monograph 2398.

MERZ, W. J., see Hooton, J. A.

MOORE, G. E., see Wooten, L. A.

MORRISON, J., and ZETTERSTROM, R. B.
Barium Getters in Carbon Monoxide, Monograph 2406.

PAUL, WILLIAM, and PEARSON, G. L.
Pressure Dependence of the Resistivity of Silicon, Monograph 2446.

PEARSON, G. L., see Paul, William

PIERCE, J. R.
Interaction of Moving Charges with Wave Circuits, Monograph 2420.

POTTER, J. F., see Fisher, J. R.

PRINCE, M. B.
Silicon Solar Energy Converters, Monograph 2419.

PURVIS, M. R., see Davey, J. R.

RAISBECK, G.
Order of Magnitude of Fourier Coefficients, Monograph 2404.

ROBERTSON, S. D.
The Ultra-Bandwidth Finline Coupler, Monograph 2464.

ROSE, D. J., and BROWN, S. C.
High-Frequency Gas Discharge Plasma in Hydrogen, Monograph 2447.

RUEHLE, A. E., see Wooten, L. A.

SCHAWLOW, A. L., see Geller, S.

TANENBAUM, M., VALDES, L. B., BUEHLER, E., and HANNAY, N. B.
Silicon n-p-n Grown Junction Transistors, Monograph 2465.

VALDES, L. B., see Tanenbaum, M

WARNER, R. see Eder, Miss M.

WATLING, R. G., see Hochgraf, L.

WEISS, M. T., and ANDERSON, P. W.
Ferromagnetic Resonance in Ferroxdure, Monograph 2431.

WILLIAMS, H. J.
Magnetic Domains, Monograph 2442.

WOOTEN, L. A., RUEHLE, A. E., and MOORE, G. E.
Evaporation of Barium and Strontium, Monograph 2386.

ZETTERSTROM, R. B., see Morrison, J.

Contributors to This Issue

MICHAEL CHRUNEY, B. S. cum laude, 1948 and M. S., 1949, Pennsylvania State University. In 1941 and 1942 he was employed by Western Electric Company. His work as a Member of the Technical Staff of Bell Telephone Laboratories dates from 1949 and includes testing and design of magnetrons, radar and switching research. He is a member of I.R.E. and a member of Eta Kappa Nu, Pi Mu Epsilon, Tau Beta Pi, Sigma Tau and Phi Epsilon Sigma.

G. C. DACEY, B.S., 1942, University of Illinois and Ph.D. 1951, California Institute of Technology. During his undergraduate days, he was employed at the Westinghouse Research Laboratories. He joined the technical staff of Bell Telephone Laboratories in 1951 and since has worked on transistor research and development. He is the author of articles for the Physical Review on transistor physics. Member of I.R.E. American Physical Society, Phi Kappa Phi, Sigma Xi, Tau Beta Pi, and Eta Kappa Nu.

GEORGE H. DOWNES, Ph.B., Sheffield Scientific School, Yale University, 1920. He has been with the Bell System since 1921, transferring with the Development and Research Department of A. T. & T. Co. to the Bell Telephone Laboratories in 1934. Throughout his career he has been associated with work in switching systems. Last year he was appointed Switching Systems Engineer with responsibility for engineering and maintenance aspects of local dial switching systems. He is a New York State Professional Engineer, a member of A.I.E.E. and Sigma Xi.

WALTER B. ELLWOOD, A.B. University of Missouri, 1924; A.M. 1926 and Ph.D. (Physics) 1933, Columbia University. His early work at Bell Telephone Laboratories from 1930 to 1935 was concerned with investigation of properties of magnetic material at very low flux densities. Later he worked on applications of magnetic material to apparatus, in the course of which he invented the glass sealed reed relay. From 1940 to 1943 he served as scientific consultant with the Bureau of Ordnance, Navy Department, Washington. He returned to the Bell Telephone Laboratories to develop pilot manufacturing processes for the reed relay.

More recently he is concerned with the fundamental physics of electrical contacts and the development of new forms of sealed switches. He is a Fellow of the American Physical Society, and the American Association for the Advancement of Science, and also a member of the Columbia University Chapter of Sigma Xi, the American Society for Metals, and the Cosmos Club, Washington.

E. L. ERWIN, in 1918, received the B.S. degree from the University of Chicago. He joined the Installation Department of the Western Electric Company in 1921, and for the next three years was occupied in installing panel offices. Following this, he joined the Technical Staff of Bell Telephone Laboratories, where he worked in the circuit laboratory until 1932. He then transferred to circuit design work and has since been engaged in development work on panel, crossbar, and PBX systems. He has been engaged in work on the No. 5 crossbar switching system since its inception.

H. W. HERMANCE joined the Laboratories in 1927 with prior experience in chemical analysis gained in toxicological and criminological work. He had also worked with the Crucible Steel Company and had spent four years with Proctor and Gamble developing analytical methods for controlling raw materials and manufactured products. During this period he carried on part time study at Newark Technical School and later at Columbia University. From 1925 to 1927 Mr. Hermance was with the Western Electric Company at Kearny working on the analytical control of materials. Since coming to the Laboratories he has specialized in developing micro-analytical methods and laboratory facilities and has had a prominent part in applying these techniques to the diagnosis of telephone manufacturing and operating problems.

M. E. HINES, B.S. in applied physics, 1940, B.S. in meteorology, 1941, M.S.E.E., 1946, California Institute of Technology. He came to Bell Telephone Laboratories in 1946, having been with the Southern California Telephone Company during his undergraduate years. His work here has been on the development of vacuum tubes for ultra high-frequency amplification and information storage. During World War II, he was a weather officer for the U. S. Air Force. He is a member of I.R.E. and Tau Beta Pi.

R. W. KETCHLEDGE, B.S., Massachusetts Institute of Technology, 1942; M.S., Massachusetts Institute of Technology, 1942; Bell Telephone Laboratories, 1942–. During World War II, Mr. Ketchledge assisted in research related to infra-red detecting devices and in the

development of sonar devices. After the war he spent two years working on the development of the Key West-Havana submarine cable system and from 1949–53 he was in charge of systems design for the L3 coaxial system. Early in 1953 Mr. Ketchledge was appointed Electronic Apparatus Development Engineer responsible for gas tube and storage tube development, and in June, 1954, he was named Switching Systems Development Engineer, responsible for major system components for electronic switching systems. Member of Sigma Xi.

G. T. KOHMAN, B.S., Kansas University, 1920; Ph.D., Yale University, 1923. Dr. Kohman joined the Western Electric Company in 1923 and Bell Telephone Laboratories in 1925. His earliest Bell System work involved studies of the absorption of water and oxygen in connection with the development of submarine cable compounds. From 1925 to 1940 he studied capacitor and other dielectric problems and was associated with the development of the dielectric used in the present central office and customer set capacitors. He was in charge of early work on the metallized paper capacitor. Dr. Kohman headed one of the groups in the wartime Manhattan Project. Since then he has worked on growth of piezoelectric crystals, development of ceramic materials, base metal contacts, dielectrics and semiconductors. At present he is in charge of the Physical Chemical Research and Development Group. Member American Chemical Society, American Ceramic Society, Professional Chapter of Alpha Chi Sigma, Tau Beta Pi, Sigma Xi, Gamma Alpha and Sigma Tau.

C. Y. LEE, B.E.E., Cornell University, 1947; M.S.E.E., 1949, and Ph.D. in Mathematics, 1954, University of Washington. John McMullen Regional Scholar, Cornell University, 1944–1947. Instructor in the Electrical Engineering Department at the University of Washington from 1948 to 1951. He joined Bell Telephone Laboratories in 1952, as a mathematician with the Switching Development Department. Member of the I.R.E., American Mathematical Society, Sigma Xi, Eta Kappa Nu and Pi Mu Epsilon.

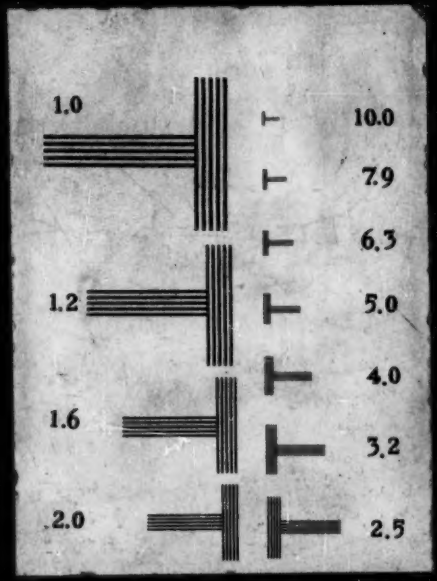
JOHN A. MCCARTHY, B.S., Union College, 1947; M.A., Columbia University, 1949; Ph.D., University of Rochester, 1952. He joined Bell Telephone Laboratories in 1954, after four years of nuclear physics research for the A.E.C. at Massachusetts Institute of Technology and the University of Rochester. He has since been concerned with design and evaluation of storage tubes. Author of two articles on radioactivity for

the Physical Review. He is a member of the American Physical Society, Sigma Xi and Phi Beta Kappa.

ARTHUR C. MEHRING, B.S.E.E., University of Maryland, 1941; M.S., Stevens Institute of Technology, 1955. A design engineer of power protective devices for Westinghouse Electric Corporation for four years, he then joined the Technical Staff of Bell Telephone Laboratories in 1945 as a member of the Switching Systems Development Department. He has been engaged principally in developing and analyzing switching circuits for the No. 5 crossbar system. For the past year he has been with the Switching Systems Engineering Department. At present he is working on the system phases of the electronic line concentrator development. Member of I. R. E., Tau Beta Pi and Phi Kappa Phi.

IAN MUNRO ROSS, B.A., Gonville and Caius College, Cambridge University, 1948; M.A. and Ph.D., Cambridge University, 1952. Dr. Ross joined Bell Telephone Laboratories in 1952, working first in transistor-physics and then in solid state device development. He is primarily concerned with the design of transistors for switching apparatus. Graduate member of the Institution of Electrical Engineers (England).

RESOLUTION CHART



100 MILLIMETERS

INSTRUCTIONS Resolution is expressed in terms of the lines per millimeter recorded by a particular film under specified conditions. Numerals in chart indicate the number of lines per millimeter in adjacent "T-shaped" groupings.

In microfilming, it is necessary to determine the reduction ratio and multiply the number of lines in the chart by this value to find the number of lines recorded by the film. As an aid in determining the reduction ratio, the line above is 100 millimeters in length. Measuring this line in the film image and dividing the length into 100 gives the reduction ratio. Example: the line is 20 mm. long in the film image, and $100/20 = 5$.

Examine "T-shaped" line groupings in the film with microscope; and note the number adjacent to finest lines recorded sharply and distinctly. Multiply this number by the reduction factor to obtain resolving power in lines per millimeter. Example: 7.9 group of lines is clearly recorded while lines in the 10.0 group are not distinctly separated. Reduction ratio is 5, and $7.9 \times 5 = 39.5$ lines per millimeter recorded satisfactorily. $10.0 \times 5 = 50$ lines per millimeter which are not recorded satisfactorily. Under the particular conditions, maximum resolution is between 39.5 and 50 lines per millimeter.

Resolution, as measured on the film, is a test of the entire photographic system, including lens, exposure, processing, and other factors. These rarely utilize maximum resolution of the film. Vibrations during exposure, lack of critical focus, and exposures yielding very dense negatives are to be avoided.